Author:
Brewer, Matt L

Title:
Investigating Fusobacterium pathogenesis using molecular and genomic methods to inform vaccine design

Investigating *Fusobacterium* pathogenesis using molecular and genomic methods to inform vaccine design

Matthew L. Brewer

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Doctor of Philosophy (PhD) in the Faculty of Life Sciences

School of Biochemistry, September 2018

Word Count: 38419

Abstract

Cellular adhesion is crucial in the pathogenesis of many bacterial species, where receptor-ligand interactions mediate host colonisation and/or invasion leading to host pathologies. *Fusobacterium nucleatum* (*Fn*) is a bacterial species gaining greater interest due to recent links with a variety of diseases, such as colorectal cancer. *Fn* is known to harbour a vast array of adhesive proteins (adhesins), however only a few have been studied in depth, such as FadA. This study aims to characterise a new previously unstudied subset of trimeric autotransporter adhesins (TAAs) found within *Fn*, responsible for binding the human receptor CEACAM1. These proteins were given the name CEACAM-binding proteins of *Fusobacterium* (CbpF).

To examine the distribution of these receptors among *Fusobacterium* spp., we screened and sequenced a library of clinical isolates. From sequencing these strains, we identified two novel species of *Fusobacterium* (*F. oralis* sp. *nov*. and *F. ovarium* sp. *nov*.), both of which harboured CbpF. While performing the taxonomic analyses on the new strains we addressed the conflicting nomenclature and phylogenetic boundaries with respect to the genus. By utilising computational methods, we could confidently delineate species and show how the genus should be organised to better reflect the genomic differences and similarities between strains.

Through screening two different types of CbpF from different species we confirmed the ability of both classes to bind CEACAM1 through proteomic- and cellular adhesion-based assays as well as showing that both classes of protein were capable of binding to CEA (CEACAM5), but not to other CEACAM variants examined. This highlighted the highly specific nature of these proteins, which was explored further by examining point mutants of CEACAM1, of which few showed any significant adhesion. As well as examining CbpF, we briefly looked at two other TAAs from *Fn*: FN0471 and FN0735; the former of which could bind indiscriminately to HeLa cells, thus indicating another important adhesin yet to be fully characterised.

Structural analysis of CbpFs highlighted a gap in the literature with respect to TAA motifs and topologies, where no known structures showed significant homology to large portions of the proteins particularly in a region predicted to be occupied with a coiled-coil motif. X-ray crystallography, SAXS and CD were used to infer structural features of CbpF, however an atomic resolution structure could not be accurately produced from a protein crystal X-ray diffraction dataset.

The work conducted here lays the foundation for additional studies into TAAs from *Fusobacterium* highlighting the requirement for increased detail on how these proteins contribute to pathogenesis and whether these proteins could be used as potential future vaccine candidates.

Acknowledgements

I would like to thank my supervisors, Dr Darryl Hill and Prof Leo Brady for providing me the opportunity to do this PhD as well as their unwavering support throughout its duration. Additionally, I would like to thank the SWBio DTP for accepting me and putting together a great doctoral training course. I also have to thank my progression panel, Ross and Paul, for providing additional support and helping to keep me on track throughout.

Secondly, I have to say a huge thank you to all the many people over the last four years that have taught me the vast array of techniques I used in this work. Firstly, to Clio for helping me massively in my early stages in the lab, as well as providing me with the wisdom of a newbie post-doc. Other members of the Hill and Brady groups that deserve my appreciation are Ahmad and Storm for showing me the ways of protein purification and to Darryl and Nibras with tissue culture and cell imaging etc. Honourable mentions go to Nick, Ash, Johnny, Sesh and various DLS staff for all their help with SAXS, CD, MD and X-ray crystallography.

In addition to the practical and academic support, I would like to thank our wonderful C-floor tech team, including Doug, Dave, Andy, Bill and Anne for all the general lab equipment support as well as making some of the less desirable buffers and media! Especially Anne for short notice media orders.

I also want to thank my office pals including Henry, Alex and various other people, including the Finn group, that have had to put up with my (and Henry's) antics over these past years. Without the relaxed working environment, I don't think I would have stayed sane over the course of my PhD.

Lastly, I would like to thank Bronwyn, my family and various other friends for being there to listen to me waffle on endlessly about a topic they clearly have no interest in and to give me a chance to not have to think about science.

Author's Declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's *Regulations and Code of Practice for Research Degree Programmes* and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

Signed:..... Date:.....

Contents

List of Figures

List of Tables

Acronyms and Abbreviations

Any acronym, initialism or abbreviation used in the main text that: belongs to the International System of Units (SI units); is a chemical element, formulae, name, nucleic acid or amino acid that conforms to the International Union of Pure and Applied Chemistry (IUPAC) nomenclature; or is the given name for a trademarked company or product, will not be listed in this table. All abbreviations are defined in the main text and this section serves as a subsidiary reference source to that.

Abbreviation	Full Name
Abs	Absorbance
Adhesin	Adhesive protein
āH	Anti-human IgG secondary antibody couple to alkaline phosphatase
Amp	Ampicillin
Amp^R	Ampicillin resistance
ANI	Average nucleotide identity
ANOVA	Analysis of variance
AP	Alkaline phosphatase
āR	Anti-rabbit IgG secondary antibody couple to alkaline phosphatase
ATCC	American Type Culture Collection
BCA	Bicinchoninic acid
BCIP	5-bromo-4-chloro-3-indolyl-phosphate
BGP	Biliary glycoprotein
BioNJ	Biological neighbour-joining
BLAST	Basic Local Alignment Search Tool
bp	Base pairs
BSA	Bovine serum albumin
CbpF	CEACAM-binding protein of *Fusobacterium*
CC	Coiled-coil
CC1	CEACAM1
CD	Circular dichroism
CD	Cluster of differentiation
CEA	Carcinoembryonic antigen
CEACAM	Carcinoembryonic antigen cell adhesion molecule
CFU	Colony forming units
Cm	Chloramphenicol
Cm^R	Chloramphenicol resistance
CRC	Colorectal cancer

CV	Column volumes
DAPI	4',6-diamidino-2-phenylindole
daTAA	Domain Annotation of Trimeric Autotransporter Adhesins
DDH	DNA-DNA hybridisation
DMEM	Dulbecco's Modified Eagle Medium
DMF	Dimethylformamide
DNA	Deoxyribonucleic acid
DPBS	Dulbecco's phosphate buffered saline
ECM	Extracellular matrix
ELISA	Enzyme-linked immunosorbent assay
Ext	Extinction coefficient
Fa	*Fusobacterium animalis*
FAA	Fastidious anaerobe agar
FAB	Fastidious anaerobe broth
FBS	Foetal bovine serum
F_C	Fragment crystallizable region of human IgG1
Fg	*Fusobacterium gonidiaformans*
Fh	*Fusobacterium hwasookii*
FITC	Fluorescein isothiocyanate
Fmal	*Fusobacterium massiliense*
Fmort	*Fusobacterium mortiferum*
Fn	*Fusobacterium nucleatum*
Fna	*Fusobacterium nucleatum subspecies animalis*
Fnec	*Fusobacterium necrophorum*
Fnecf	*Fusobacterium necrophorum subspecies funduliforme*
Fnecn	*Fusobacterium necrophorum subspecies necrophorum*
Fnn	*Fusobacterium nucleatum subspecies nucleatum*
Fnp	*Fusobacterium nucleatum subspecies polymorphum*
Fnv	*Fusobacterium nucleatum subspecies vincentii*
Fnw	*Fusobacterium nucleatum subspecies W1481*
For	*Fusobacterium oralis sp. nov.*
Fov	*Fusobacterium ovarium sp. nov.*
Fperf	*Fusobacterium perfoetens*
Fperio	*Fusobacterium periodonticum*
FPLC	Fast protein liquid chromatography
Fpoly	*Fusobacterium polymorphum*
Fpperio	*Fusobacterium pseudoperiodonticum sp. nov.*
Fpvar	*Fusobacterium pseudovarium sp. nov.*
Fr	*Fusobacterium russii*
Fu	*Fusobacterium ulcerans*
Fv	*Fusobacterium vincentii*

Fvar	*Fusobacterium varium*
Fw	*Fusobacterium W1481*
g	Acceleration as a factor of gravity
GOI	Gene of interest
GPI	Glycosylphosphatidylinositol
GPU	General-purpose graphics processing unit
HPC	High performance computing
HPLC	High performance liquid chromatography
hr	Hours
IBD	Inflammatory Bowel Disease
Ig	Immunoglobulin
IPTG	Isopropyl β-D-1-thiogalactopyranoside
ITAM	Immunoreceptor tyrosine-based activating motif
ITIM	Immunoreceptor tyrosine-based inhibitory motif
ITS-G	Gibco® Insulin-Transferrin-Selenium
Kan	Kanamycin
Kan^R	Kanamycin resistance
LB	Luria-Bertani
LC-MS	Liquid chromatography-mass spectroscopy
L-Glu	L-Glutamine
LIC	Ligation-independent cloning
LPS	Lipopolysaccharide
Mbp	Mega base pairs
MBP	Maltose-binding protein
mCEACAM	Murine carcinoembryonic antigen cell adhesion molecule
MD	Molecular dynamics
min	Minutes
MLSA	Multi-locus sequence analysis
MOI	Multiplicity of infection
Mr	Relative molar mass
MUM	Maximal unique matches
MUMi	Maximal unique matches index
MWCO	Molecular weight cut-off
NBT	4-nitro blue tetrazolium chloride
NCBI	National Centre for Biotechnology Information
NCTC	The National Collection of Type Cultures
NEB	New England Biolabs
NJ	Neighbour-joining
NK cells	Natural Killer cells
NTA	Nitrilotriacetic acid
O/N	Overnight

OD	Optical density
OMP	Outer-membrane protein
PBC	Periodic boundary conditions
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
PDB	Protein Data Bank
Pen	Penicillin
PFA	Paraformaldehyde
PLSA	Pan-locus sequence analysis
PPMCC	Pearson's Product Moment Correlation Coefficient
PSG	Pregnancy-specific glycoprotein
rDNA	Ribosomal deoxyribonucleic acid
R_g	Radius of gyration
RMSD	Root-mean-square deviation
RNA	Ribonucleic acid
RPMI-1640	Roswell Park Memorial Institute medium 1640
rRNA	Ribosomal ribonucleic acid
RT	Room temperature
SAXS	Small-angle X-ray scattering
sec	Seconds
SEC	Size-exclusion chromatography
Sigma	Sigma-Aldrich
SOC	Super Optimal broth with Catabolite repression
sp.	Species (singular)
sp. nov.	*Species nova*
spp.	Species (plural)
spp. nov.	*Species novae*
Strep	Streptomycin
subsp.	Subspecies
T5cSS	Type Vc Secretion System
T5SS	Type V Secretion System
TAA	Trimeric autotransporter adhesin
TBE	Tris-Borate-EDTA
Thermo	Thermo Fisher Scientific
TIGIT	T cell immunoreceptor with Ig and ITIM domains
UPGMA	Unweighted Pair Group Method with Arithmetic Mean
Usp	Ubiquitous surface protein
v/v	Percent volume in a volume
VE-cadherin	Vascular endothelial cadherin
w/v	Percent weight in a volume
WGS	Whole-genome sequencing

X-Gal	5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside
Zeo	Zeocin
Zeo^R	Zeocin resistance

Chapter 1: Introduction

Cellular adhesion plays an important role in pathogenesis for many bacterial pathogens, where receptor-ligand interactions facilitate a broad range of functions, for example, in mediating host colonisation and/or invasion potentially leading to disease. *Fusobacterium nucleatum* (*Fn*) represents a bacterial species gaining greater interest due to recent links with a variety of pathologies, such as colorectal cancer (1-3). *Fn* is known to harbour a vast array of adhesive proteins (adhesins) (4, 5), however very little is known regarding their roles in pathogenesis with only a few being studied in depth, such as FadA (6). This study aims to characterise a new previously unstudied subset of adhesins found within *Fn*.

1.1 The *Fusobacterium* genus

1.1.1 Biophysical and Biochemical Characteristics

Fusobacterium comprises a heterogeneous bacterial genus, however each species shares some fundamental traits. *Fusobacterium* species are obligate anaerobic, Gram negative, nonsporulating, nonmotile bacteria that adopt a pleomorphic filamentous shape; hence the name originates from the Latin *fusus* meaning spindle (7, 8). The lengths of these bacteria can range from 1 to 10 μm making them some of the longest bacteria (**FIGURE 1.1**).

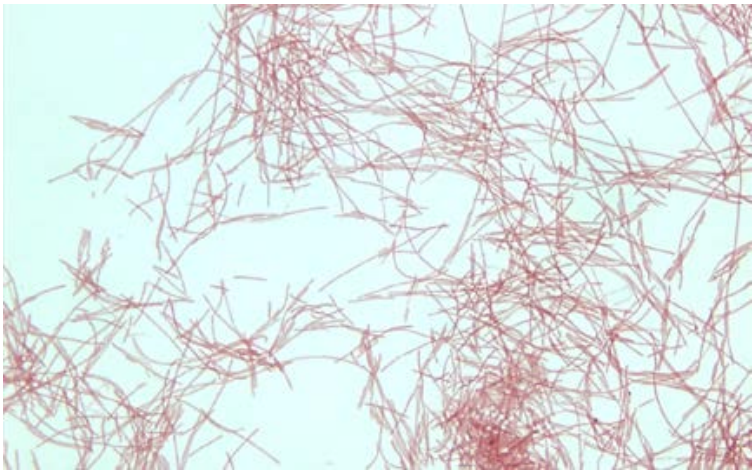


Figure 1.1 | **Gram-stained *Fusobacterium nucleatum* micrograph.**

Image obtained from Partners' Infectious Disease Images eMicrobes Digital Library (https://www.idimages.org/images/detail/?imageid=943).

Fusobacterium spp. ferment carbohydrates yielding butyric acid as the primary product and will only grow in atmospheres containing typically less than 5 % O_2 though they can tolerate up to 6 % (9). In addition to using glucose and peptones as metabolites, *Fusobacterium* will also use amino acids such as lysine, glutamate, aspartate and histidine, which is contrary to most other rod-shaped Gram-negative bacteria, though it has been shown that *Fusobacterium* will bias the use of peptides over free amino acids (10-12). A further unusual property was discovered where it was found glucose is not the primary metabolite for energy production and is, however, used for the synthesis of other small molecules within the cell (13-16). It has also been demonstrated that glutamate alone can be used as the primary energy source for *Fn* (17).

1.1.2 Taxonomy

The *Fusobacterium* genus exists within the *Fusobacteriaceae* family and form the predominant group of bacteria that have been characterised, where they are subdivided into several distinct species. Of these species, arguably the most studied is *F. nucleatum* (*Fn*). This species is very diverse and as such, was historically split into several distinct subspecies. The currently acknowledged subspecies are: *nucleatum* (*Fnn*), *vincentii* (*Fnv*), *animalis* (*Fna*), *polymorphum* (*Fnp*) and W1481 (*Fnw*). These classifications are a conglomeration of multiple previous literature (18-21). As will be explored later and which has been discussed in recent studies (22, 23), these boundaries may not be definitive where the reorganisation of the genus may be required. For example, for a brief period, another subspecies called *fusiforme* (19) existed, though it has since been disregarded due to its overwhelming likeness to *Fnv* when comparing whole-genome data (20).

F. necrophorum (*Fnec*) accounts for the majority of other research within the genus. Like *Fn*, *Fnec* has been categorised into subspecies: *Fnec* subsp. *funduliforme* (*Fnecf*) and *Fnec* subsp. *necrophorum* (*Fnecn*), where the names originate from previously misclassified bacteria (24). With respect to health and disease, this species has the most far-reaching documented consequences, affecting both animals and humans (24).

In addition to these, there are also other minor and uncategorised species with comparatively little information. On the same major evolutionary lineage as *Fn*, is *F. periodonticum* (*Fperio*) and the newly discovered *F. hwasookii* (*Fh*), though fewer studies have been undertaken and both have so far only been associated with periodontal disease (25, 26). The current list of other minor strains with whole-genome sequencing (WGS) data, at the time of writing, are as follows: *F. equinum*, *F. gonidiaformans*, *F. massiliense*, *F. mortiferum*, *F. naviforme, F. perfoetens*, *F. russii*, *F. ulcerans* and *F. varium*. These strains have been associated with a diverse set of diseases, but a lack prevalence in humans is the likely cause for there to be less information on them. Other strains such as *F. necrogenes*, *F. simiae* and *F. canifelinum* have been identified using 16S rRNA sequencing, but their genomes have yet to be sequenced and therefore will not be included herein. The evolutionary relationship, as determined by 16S rRNA comparisons, within the *Fusobacterium* genus is displayed in **FIGURE 1.2**.

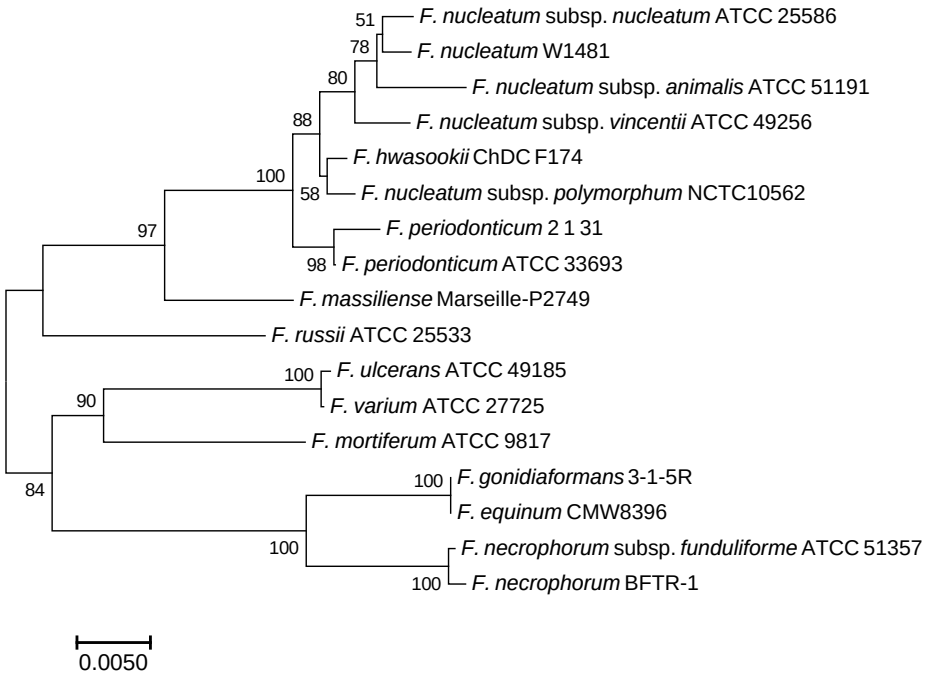


Figure 1.2 | **The *Fusobacterium* genus organisation.**

This dendrogram shows the evolutionary history between representative strains from *Fusobacterium* as determined by their 16S rRNA sequence.

1.1.3 Health and Disease

1.1.3.1 Periodontal Disease

F. nucleatum primarily resides in the oral cavity, where it is often located within dental plaque. *Fn* is particularly promiscuous and can act as a 'biological cement' via coaggregation with many other organisms forming biofilms. It relies on this nature to colonise teeth, as it cannot bind enamel directly and so binds to the existing primary colonisers (27, 28). With poor oral hygiene, over time, more pathogenic organisms that cannot bind to primary colonisers or tooth enamel directly, such as *Porphyromonas gingivalis* and *Aggregatibacter actinomycetemcomitans*, have a chance to establish themselves utilising *Fn* as a scaffold (**FIGURE 1.3**).

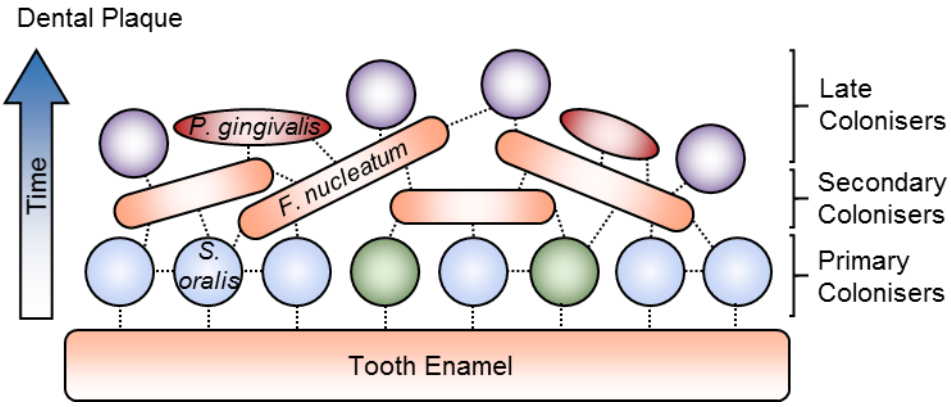


Figure 1.3 | **Dental plaque formation over time.**

Primary colonisers such as *S. oralis* and *S. sanguinis* can bind directly to the tooth surface and hence will colonise the area quickly. The majority of bacteria are unable to colonise directly, including *Fn*, where they colonise using interactions with the primary colonisers. Once *Fn* has colonised, this then allows for the establishment of a wide array of other organisms via interactions with *Fn*. Some of these late colonisers can be more pathogenic, such as *P. gingivalis*, and lead to more severe diseases.

These later colonisers, with *Fn*, can then go on to invade the subgingival crevice causing periodontitis. *Fn* thrives here as this environment is much lower in oxygen (27, 29, 30). **FIGURE 1.4** shows how periodontitis compares to a healthy state.

An inability of the immune system to clear the infection, or the lack of appropriate treatment, consequently leads to chronic infection, where the production of acidic by-products from the invading organisms can slowly dissolve the tooth and the release of inflammatory cytokines by the immune system can trigger immune-mediated damage to the surrounding tissues. If left untreated, as well as localised damage, further complications can arise due to gingival bleeding when tooth brushing, allowing for bacteria to enter the bloodstream leading to a state of transient bacteraemia, which can go on to cause other diseases such as bacterial endocarditis (31).

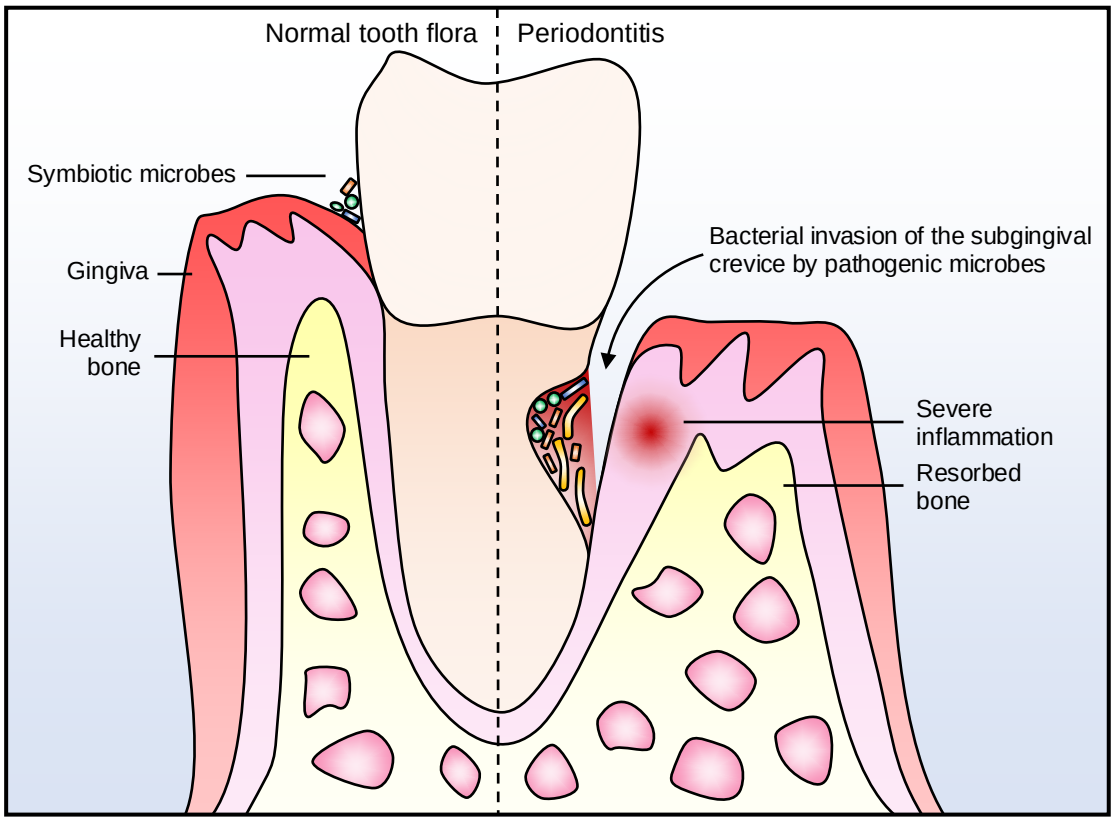


Figure 1.4 | **Periodontal health compared with periodontitis.**

The left shows the state of periodontal health where only resident commensals microbes are present. The right shows advanced periodontitis where *Fn* and late colonisers invade the subgingival crevice causing severe inflammation and tissue degradation. Adapted from Hajishengallis *et al*. Nat Rev Immunol. 2015 (32).

1.1.3.2 Colorectal cancer

In recent years, *Fusobacterium* has been linked with colorectal cancer (CRC) progression, where numerous studies have shown significantly increased levels of the bacteria in colorectal adenomas and carcinomas (33-40). As such, it can be determined that there is a definite association, however, there remains debate over what the consequences and importance of this are. One study in mice showed direct evidence for the presence of *Fusobacterium* leading to more rapid tumourigenesis and poorer prognoses (41), however this study does not provide evidence for why this occurs and may not give an accurate model for the human case, especially when other environmental factors come into play. Other studies have provided evidence for the role of *Fusobacterium* only being of importance in the later stages (stage III to IV) of CRC progression and not in the initial stages (42, 43).

Various models have been proposed for the activity of certain virulence factors within *Fusobacterium*. Two outer-membrane proteins (OMPs) have been indicated as potential aetiological agents in CRC: FadA (6, 44) and Fap2 (45). FadA is responsible for binding to vascular endothelial cadherin (VE-cadherin) which can the mediate β-catenin signalling in the cell and could lead to the generation of a proinflammatory microenvironment (44, 46). Fap2 instead may aid in creating an immunosuppressive microenvironment via binding and signalling through the TIGIT (T cell immunoreceptor with Ig and ITIM domains) receptor on Natural Killer (NK) cells (45). It remains unclear what these two proteins may be doing *in vivo*, however they have both been shown they have potential to be involved in disease.

Other studies have examined the origin of the *Fusobacterium* species that have been isolated from CRC and there is evidence to suggest that the bacteria associated with CRC adenomas are haematogenous and have not traversed from the colon lumen. Moreover, it is unclear if the bacteria travel with metastases or if they reconvene after establishment at a secondary site (47, 48).

Some research has suggested that the association could be entirely circumstantial where the increased expression of CRC biomarkers leads to increased adhesion of these bacteria, nevertheless, further research is required in this field to determine the true impact of the bacteria on this disease. Bashir *et al* and Han provide more in-depth reviews on the topic (49, 50). In addition to colorectal cancer, *Fusobacterium* species have been linked with pancreatic and oral cancers, which have yet to be as thoroughly explored as CRC (51, 52).

1.1.3.3 Preterm births

In periodontal disease, *Fn* itself has a chance enter the bloodstream from gingival bleeding, giving rise to a state of transient bacteraemia. This is of particular issue in pregnant women; it is known that poor oral hygiene during pregnancy is a risk factor for adverse foetal development and even miscarriage (53-55). The mechanisms of this are not yet fully characterised, though *Fn* species are the most commonly isolated bacteria from significant tissues such as the placenta from preterm births. Interestingly, the only two subspecies that have been isolated from intrauterine infections are *Fna* and *Fnp* (56-58).

The FadA protein has been indicated, again, in the model for infection upon colonisation of the placenta. It was demonstrated that a *fadA* knockout mutant was unable to bind the murine placenta where the re-complemented mutant restored binding (59, 60). It is proposed that the redistribution of VE-cadherin upon binding by FadA increases permeability of the tissue disrupting tight junctions, therefore allowing for bacteria to traverse the barrier (44).

1.1.3.4 Lemierre's Syndrome

Currently, the only human disease attributed almost exclusively to *Fusobacterium* spp. is Lemierre's Syndrome. This disease is most commonly characterised by thrombophlebitis of the internal jugular vein when peri-tonsillar abscesses containing *Fusobacterium* rupture and leak into the vein (61). It is often fatal if left undiagnosed and largely affects young healthy individuals (61). Other sites have been shown to be affected such as the hepatic portal vein (62). Emboli can break off the initial site of infection and travel through the

bloodstream leading to coronary or pulmonary embolisms for example, therefore patient fitness becomes largely irrelevant at this point, thus the indiscriminate relationship to age.

Fnec species, specifically *Fnec* subsp. *funduliforme*, cause approximately 80 % of all diagnosed cases and a further 10 % caused by other *Fusobacterium* species (61, 63). Unfortunately, from a patient's and a researcher's perspective, this disease is both rare (~1 per million incidence) and commonly misdiagnosed (61, 64). This makes identifying risk factors and bacterial relationships with this disease hard to study. A typical case can present with a sore throat and general malaise, so combined with its rarity and the reduction in prescribing antibiotics for sore throats, it is no surprise this disease is being mistreated and incidence may even be increasing as a result (65).

1.1.3.5 Footrot

Footrot is a disease restricted to hooved animals such as sheep and cattle that inflicts a particularly heavy burden on sheep farmers. *F. necrophorum* is one of two causative agents of infection, the other being *Dichelobacter nodosus*, where coinfection of the two is required for disease (66, 67). The disease is highly contagious between animals, though it is thought not to transmit between species, such as sheep and cattle (67). Due to the high transmission rate between animals, whole herds can become infected if the diseased animal is not quarantined quickly. As a result, this disease has large financial implications, for example, the estimated cost of ovine footrot to UK farmers alone was £24.4 million in 2005 (68).

The disease is characterised by the degradation of the interdigital tissue between the toes of the animal caused by keratinases and proteases produced by *D. nodosus* (67). This can lead to lameness of the affected animal as it becomes painful to stand due to the infection. Treatment with antibiotics should clear the infection and limit further damage, though there have been cases reported with infection that develops much faster and cannot be dealt with using normal treatment, often leading to euthanising the animal. Some vaccines have been marketed, though the effectiveness has been doubted and they only offer a limited immune duration (69).

1.1.3.6 Other Diseases

In addition to the listed pathologies, *Fusobacterium* spp. have been shown to be implicated in several other diseases. These cases are generally less numerous and not as well characterised but interesting nevertheless. Many of these other associated diseases are thought to be onset from a prior case of periodontitis. For instance, *Fn* were detected within the synovial fluid of an arthritic joint where the exact same bacterial clone was also identified in a case of periodontitis within the same patient (70). Other diseases *Fusobacterium* spp. has been associated with include (but are not limited to): inflammatory bowel disease (71, 72), atherosclerosis (73), appendicitis (74) and it has even been indirectly implicated in Alzheimer's disease (75).

1.1.3.7 Treatment

The primary method for treating *Fusobacterium*-related infections rely upon physical removal of the biofilms (in dental applications) and using broad-spectrum antibiotics, such as penicillin or cephalosporin derivatives, or tetracycline (76). These are usually accompanied by an additional anaerobic bacterium-targeting drug such as metronidazole, as beta lactamases have since been identified within some strains of *Fusobacterium*. However, metronidazole can cause some severe side-effects so is not always used (77). More recently, the anaerobe-targeting drug Amixicile could be a possible alternative treatment as it selectively targets anaerobic organisms via the pyruvate:ferredoxin oxidoreductase enzyme (78). This drug would also solve the problem of potential onset of *C. difficile* colitis that treatment which metronidazole or clindamycin could cause (79).

1.1.4 Virulence Factors

A number of proteins and cellular components have been implicated as aetiological agents in *Fusobacterium*-related diseases.

Extracellular vesicles derived from *Fusobacterium* cells have been shown to co-aggregate with *P. gingivalis* and have been shown to possess proteolytic activity, which could be important in various pathological pathways, such as tissue invasion and immune

sequestering (80, 81). A serine protease found on the surface of these vesicles, named fusolisin, was found to degrade extracellular-matrix (ECM) proteins fibronectin, fibrinogen, and collagen I and IV. This protease also showed the ability to degrade IgA alpha-chains. The ability to degrade ECM components highlights a role for this protein in tissue invasion (81, 82).

Secreted haemolysin is also considered to be a major virulence factor in *Fusobacterial* pathogenesis whereby lysis or erythrocytes yields iron and can lead to a hypoxic environment at the site of infection, which is favourable for the obligate anaerobic nature of *Fusobacterium* spp. (83, 84).

1.1.4.1 Adhesive Proteins

Crucial to *Fusobacterium* pathogenesis are its vast array of adhesins. FadA is likely the most studied of the adhesins that some *Fusobacterium* strains possess (6). This protein is relatively small at around 150 residues long and exists within the outer membrane in *Fnn*, *Fnv*, *Fna* and *Fnp*. As previously stated, it is responsible for binding to cadherins on human cells, specifically VE-cadherin (44). In addition to this receptor, other adhesins have been identified as key factors in colonisation and disease.

The major outer-membrane protein, FomA plays a key role in biofilm formation as well as serving other functions for *Fusobacterium*. It was identified as a voltage-dependant porin that was capable of binding to F_C (Fragment crystallizable region) domain of human IgG which aids evasion of the immune system by sequestering potentially immune-activating antibodies (85-87). This protein also has the property of binding other bacteria such as *P. gingivalis* and *S. sanguinis*, therefore facilitating a bridge between commensal and pathogenic organisms (80).

In addition to adhesion to TIGIT, as previously mentioned, the Fap2 protein is also involved in other interspecies interactions, specifically to *P. gingivalis* (88). Likewise, the OMP RadD, a Type Va secretion system, has been shown to be involved in enabling interactions with

P. gingivalis as well as various less pathogenic Gram-positive bacteria such as *A. naeslundii*, *S. gordonii*, *S. mutans* and *S. oralis* similarly to FomA playing a role in bacterial coaggregation (89, 90). Recently, another coaggregation-promoting OMP was identified in *Fnn* called CmpA, which was shown to allow dual-species biofilm formation with *S. gordonii* (91).

In addition to these proteins, in *Fnn* three genes encoding putative Type Vc Secretion Systems have been identified. The gene loci for these are FN0471, FN0735 and FN1499 (*Fnn* ATCC 25586). Hitherto, none of these genes have been examined and are likely facilitating adhesion to unidentified targets. It is products from these genes, and related proteins, that this study aims to characterise.

1.2 Trimeric Autotransporter Adhesins

Trimeric autotransporter adhesins (TAAs), also known as Type Vc Secretion Systems, encompass a large variety of bacterial outer-membrane proteins. These proteins are used mainly for facilitating adhesion between a bacterium and substratum, such as human or other bacterial cells (92). They are close relatives of the classical Type V Secretion System (T5SS) where several key differences exist, which will be explained.

1.2.1 Structure and Assembly

The structures of TAAs varies widely, however all contain a few common features. The first is a signal peptide that, after initial folding in bacterial cytoplasm, will traffic the proteins to the inner-membrane where they cross into the periplasm via the Sec pathway. Here the signal sequence will be cleaved, and three protein monomers will form a trimer before inserting into the outer membrane (92-94). This is very similar to how the classical Type V secretion pathway works; however, the protein does not undergo cleavage when inserted into the membrane and instead retains the extracellular domain, also known as the passenger domain, unlike the classical T5SS which normally secretes the passenger domain (95).

The most conserved region of TAAs is the membrane anchor – this serves to hold the protein in the outer membrane of the bacterium, as well as serving to export the rest of the protein to the extracellular space on construction (92, 93). The membrane anchor is found at the C-terminus of the protein and can sometimes have extra amino acid residues further downstream that reside in the intermembrane space. At the core of this domain is a homotrimeric 12-stranded β-barrel fold containing a coiled-coil that traverses the length of the β-barrel and links to the extracellular domain. The fold is highly conserved among TAAs with little to no variation found. **FIGURE 1.5** displays the membrane-spanning structure of Hia, a TAA from *Haemophilus influenzae*.

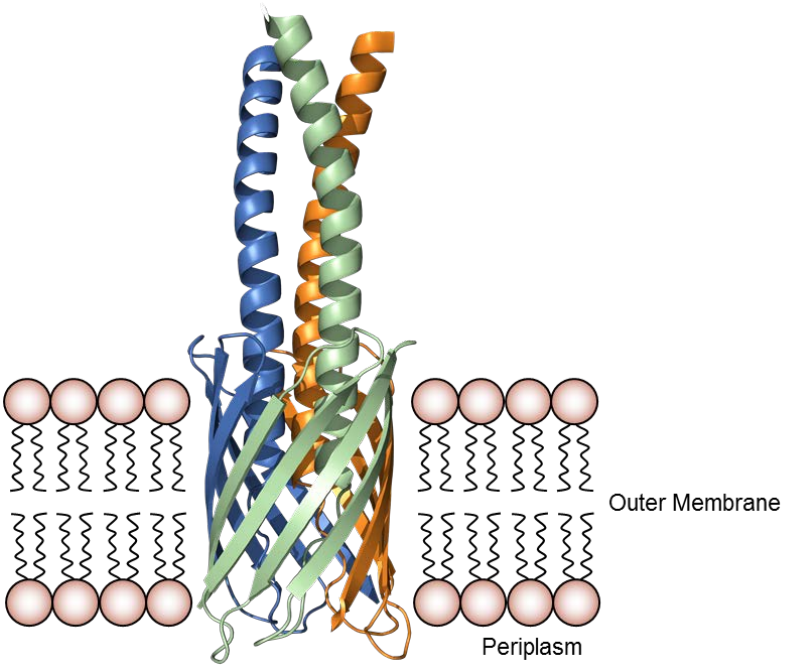


Figure 1.5 | **Membrane anchor structure for a trimeric autotransporter adhesin.**

The structure of Hia [*H. influenzae*] residues 992-1098 is shown (PDB ID: 2GR7) (96) and its state post outer-membrane insertion. The membrane anchor is comprised of a 12-stranded β-barrel composed of three identical monomers (3 x 4 β-sheets). The length of the coiled-coil extruding from the membrane anchor varies in length depending on the protein.

1.2.2 Specialised folds

The structure of the region flanked by the membrane anchor and the signal peptide can contain a wide variety of different folds that can be split into three classes: heads, connectors and stalks (93, 97).

Head domains contain predominantly β-sheet motifs. The head class can be split into two subclasses: transverse and interleaved heads, where transverse heads have their β-strands perpendicular to the longitudinal axis and interleaved are parallel. Transverse head domains are likely to contain repeating sequential runs of a particular motif varying in length, whereas interleaved do not repeat and have a set size (97, 98). One common fold in the transverse head class is represented by the architype TAA, YadA, from the bacteria *Yersinia enterocolitica*. This structure, known as the YadA-like head domain has a characteristic β-solenoid motif which can repeat any number of times giving rise to multitude of head lengths (97, 98). This structure is probably the most common head domain type found within TAAs. This structure provides an extracellular trimerization region where the head regions are held tightly to each other through a hydrophobic core (**FIGURE 1.6**) (99, 100).

The number of residues per turn of YadA-like heads varies between individual proteins but is usually 14 or 15 residues for TAAs. Kajava & Steven (101) provide a detailed look at other β-solenoid motifs from Type V Secretion Systems and how the equivalent domains differ between the classical T5SS and TAAs (T5cSS).

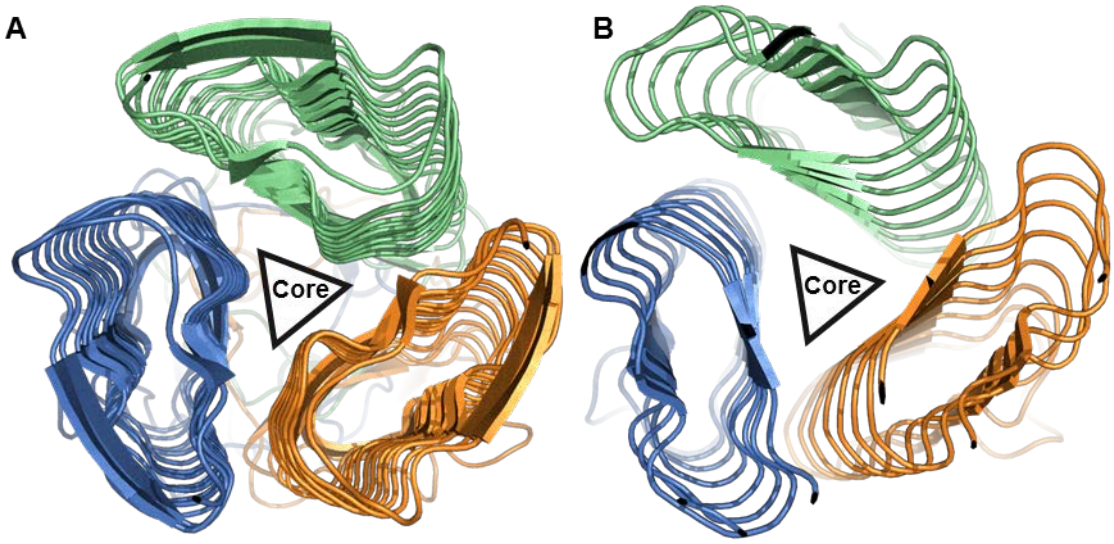


Figure 1.6 | **YadA-like Head Motif.**

The structure of the YadA-like head domains from the partial structures of **A)** UspA1 (PDB ID: 3PR7) (102) and **B)** YadA (PDB ID: 1P9H) (103). This fold adopts an O-shaped beta roll, also known as the β-solenoid motif, with 15 and 14 residues per turn for UspA1 and YadA respectively. The number of residues per turn can vary between alternate TAAs. The core of the head domain is stabilised by hydrophobic residues and prevents the head domain monomers from dispersing.

An example of an interleaved head domain is the Tryptophan-Ring which is made up of a β-meander motif. It is the most common variant of interleaved heads and others identified all use this as a backbone structure. An example of this fold is shown in **FIGURE 1.7**. This structure can be found in the protein BadA from *B. henselae*. The head domain of this protein was shown to be essential for host cell adhesion and interactions (104).

Stalks form the next common feature of TAAs. The stalk domains of TAAs are usually made up of trimeric coiled-coils, which can adopt either a right-handed or left-handed shape and may also contain various inserts. **FIGURE 1.8** shows a typical arrangement of a coiled-coil. TAA stalks can adopt several amino acid motifs for residues $a - g$, an example heptameric peptide that some stalks with a polar core use is: LXXTNXN (where X is undefined), though more commonly small hydrophobic residues are found in the d register (98).

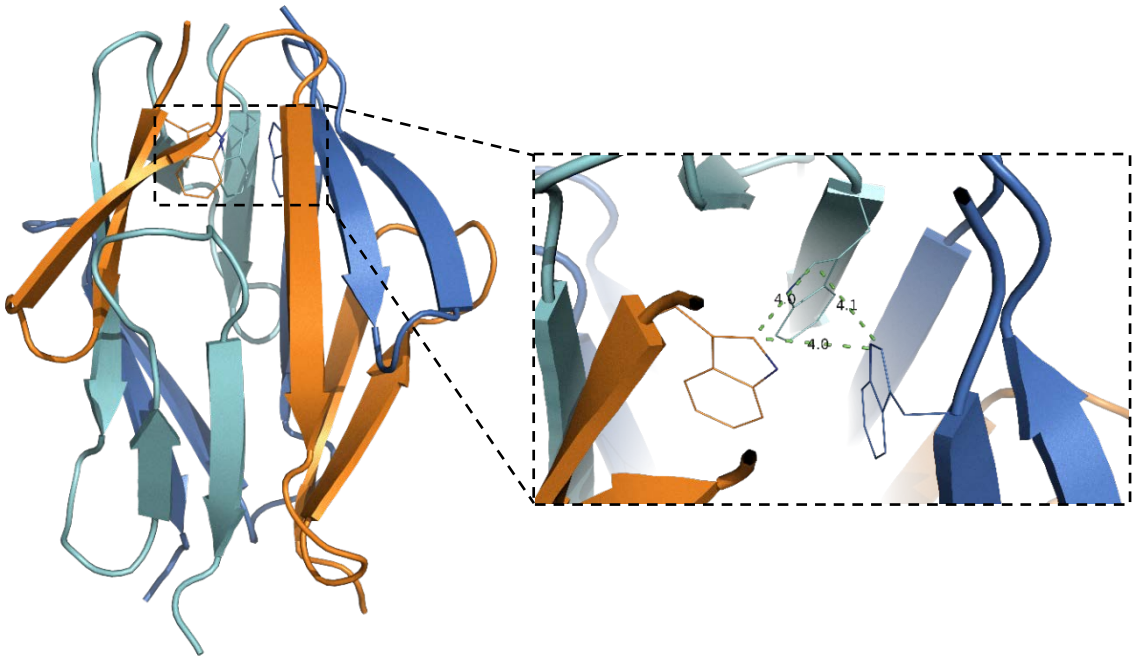


Figure 1.7 | **The Tryptophan Ring Motif.**

This is the structure of the Trp-Ring β-meander fold of BadA (residues 385-431) from *Bartonella henselae*. This is an example of an interleaved head motif. The Trp-Ring head is characterised by the interactions between three tryptophan residues, one from each chain spaced approximately 4 Å apart. Each chain of the trimer is coloured for clarity.

Some TAAs can incorporate both a left-handed and right-handed CC into their structure such as the IgG-binding protein EibD from *E. coli* (105). Interestingly, in this case, it utilises the right-handed CC to bind human IgA and the left-handed CC to bind human IgG. Connecting the two opposite-handed coiled-coils is a domain known as the Eib saddle, which is composed of a short loop.

The final common TAA domain are the necks which are responsible for connecting head and stalk domains. These can vary in length from 19 to 22 residues long but show higher variation between different TAAs (98). A detailed list of known TAA domains is given in **TABLE 1.1**.

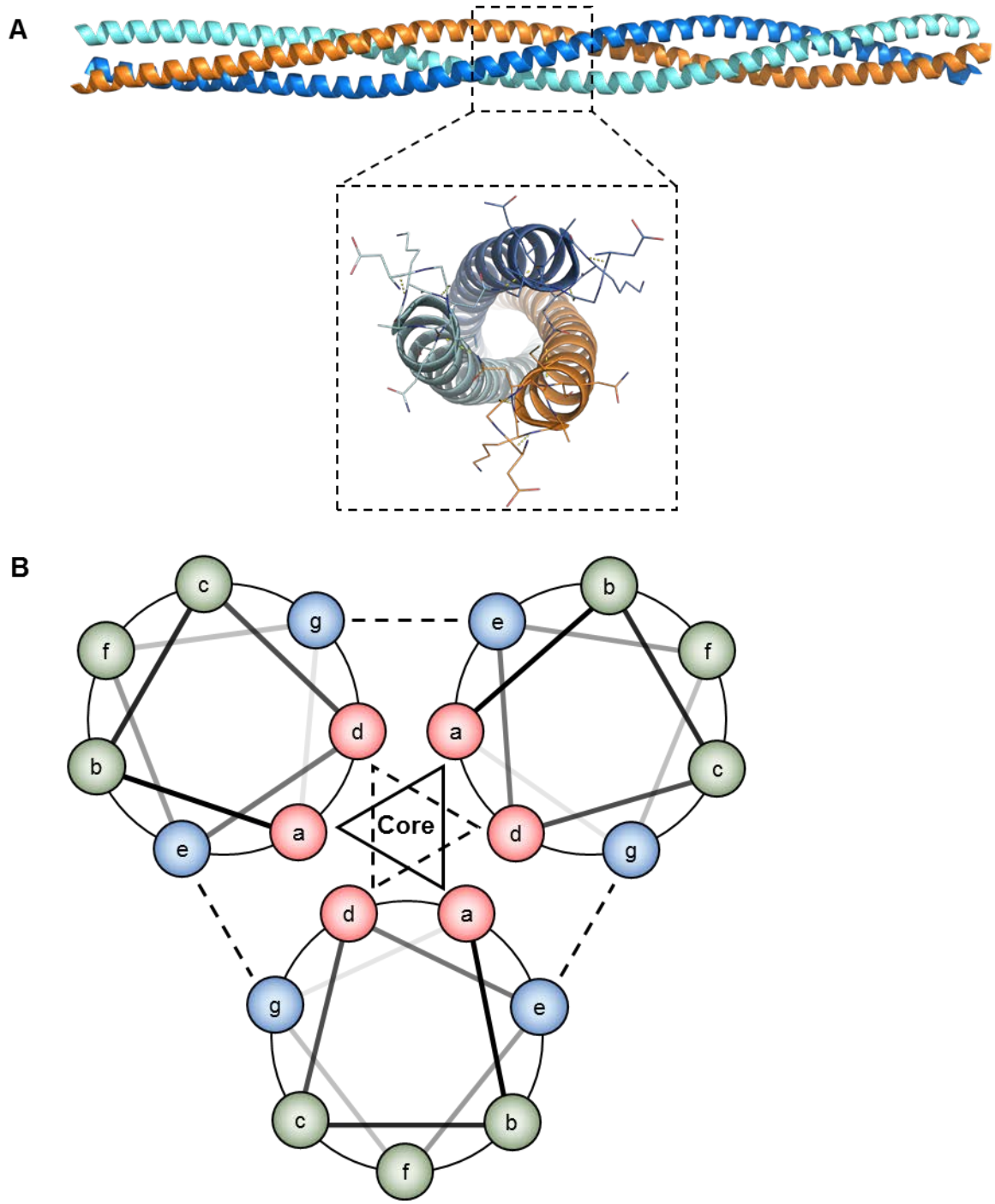


Figure 1.8 | **Trimeric Coiled-coil Structure.**

A) The atomic structure of the UspA1 stalk fragment (amino acids 527-665), rD-7. Seven amino acids from each chain have been shown as lines to demonstrate one repeating unit. **B)** Amino acid residue positions within coiled-coils. Letters $a-g$ represent a heptameric peptide repeat as part of an alpha helix. Residues at positions a and d face inwards towards the core and contribute most to structure formation through interactions $a-a-a$ and $d-d-d$. Residues at e and g can further stabilise the coiled-coil using polar interactions.

Table 1.1 | **Known common domains and motifs of TAAs.**

A brief description of each domain is given, stating its class, structure and function (if applicable). In addition, example structures of each domain are given if known or relevant. [1] position d refers to the position of an amino acid residue on a coiled-coil structure, see **FIGURE 1.8-B**. [2] Hia [*H. influenzae*]. [3] YadA [*Y. entercolitica*]. [4] BpaA [*B. psudomallei*]. [5] UspA1 [*M. catarrhalis*]. [6] SadA [*S. enterica*]. [7] BadA [*B. henselae*]. [8] EibD [*E. coli*]. *Table adapted from tables in* (97, 98).

Domain	Description	Example PDB ID	Reference(s)
Membrane anchor	12-stranded β-barrel	2GR7[2]	(96)
Signal peptide	Responsible for trafficking pre-protein to periplasm, cleaved after this	-	(106)
Ylhead	Transverse head, β-solenoid motif	1P9H[3]	(103)
HIM1, 2 & 3	Insert in final Ylhead repeat prior to the neck	3LAA[4], 3PR7[5] & 2YO0[6]	(102, 107, 108)
GIN	Transverse head, β-prism motif, found only after interleaved heads	3D9X[7]	(109)
TrpRing	Interleaved head, β-meander motif, most common of its type	3D9X	(109)
FxG	Interleaved head, variant of TrpRing	-	
GANG	Interleaved head, deletion variant of TrpRing	-	
Stalks	α-helices in trimeric left or right handed coiled-coil (CC)	1P9H	(103)
N@d^1	Asn residue in position d of CC, most common polar amino acid in the 'core' of TAA CCs.	-	(110)
FGG	3-stranded β-meander insertion into CC	2YO2[6]	(108)
Eib saddle	Unique to Eib proteins, insertion of non-helical structure into CC	2XQH[8]	(105)
Neck	Connector, links β-sheet motifs to α-helices	1P9H	(103)
Short neck	Connector, neck variant containing 19 amino acids	3D9X	(109)
Long neck	Connector, neck variant containing 22 amino acids	3LAA	(107)
KG	Connector, neck variant missing initial β-strand	3EMI[2]	(111)
DALL	Connector, links α-helices to β-sheets, exclusively found before neck domains, three known conserved variants (DALL1-3)	2YO3[6]	(108)
HANS	Connector, links β-sheet motifs to α-helices, always followed by a Ylhead	2YO3	(108)

1.2.3 Receptor binding

TAAs are used by bacteria to bind to many different structures, such as human surface receptor proteins or ECM proteins. It has been found that the presence of these adhesins can dramatically increase virulence in some pathogens, such as BpaB in *Burkholderia mallei* (112), indicating their adhesive roles are sometimes necessary to cause disease.

The protein BadA from *B. henselae* is a TAA that uses its different domains to selectively bind certain ECM components. It utilises both its head and stalk domains to bind to collagen in a redundant fashion whereas binding to fibronectin is exclusively mediated by the stalk domain (113).

The ubiquitous surface proteins (Usp) A1 and A2 proteins from *Moraxella catarrhalis* bind to several different receptors, each having different roles in pathogenesis. UspA1 can adhere to the human receptor carcinoembryonic antigen cell-adhesion molecule 1 (CEACAM1) on epithelial cells in the respiratory tract, aiding establishment of the bacteria in the host (114). In addition, these proteins interact with extracellular matrix proteins, such as laminin, fibronectin and vitronectin (115-117). UspA2 binding to vitronectin has been demonstrated to be important in immune escape by interfering with complement-mediated killing (117).

The CEACAM1 binding region in UspA1 has been determined to be the coiled-coil stalk domain. This region has been isolated and expressed as the recombinant peptide rD-7 (118), which was shown to bind CEACAM1 specifically. This is another occurrence of a TAA using its stalk domain to bind an Ig-like protein, in addition to EibD (105).

Through an unpublished study, it was found that the FN1499 protein from *Fn* coprecipitated with CEACAM1. This was identified through sequencing of the N-terminal domain using Edman degradation. Similarly to UspA1, this is another identified TAA protein responsible for cellular adhesion using the CEACAM1 receptor. These proteins have been labelled CbpF for CEACAM-binding proteins of *Fusobacterium*.

1.3 CEACAMs

The carcinoembryonic antigen (CEA) family of proteins are members of the immunoglobulin superfamily that contains several cell surface proteins that are expressed variably on different tissues throughout the body (119). The family is split into two branches: CEA cell adhesion molecules (CEACAMs) and pregnancy specific glycoproteins (PSGs) (120). CEACAMs are thought to be evolutionarily recent, mainly appearing in mammals, though orthologues in other species have been identified (120-122). They can interact with each other in either a homo- or heterotypic fashion and are involved in cell-cell adhesion, but also have many other functions (123-126).

CEACAMs come in many different types, each having a different structure with varying features (**TABLE 1.2**). The genes encoding CEACAMs can also produce altering products due to exonic splice events, for example, the gene for CEACAM1 (also known as CD66a and BGP1) has 12 known protein products (120). Some of the resulting splice variants from CEACAMs may result in a truncated extracellular domain or produce soluble protein with no membrane anchor which can then be secreted by the cell.

1.3.1 Functions

Some of the features CEACAMs can contain, relate to their functions, for example, CEACAM1 and 3 have immunoreceptor tyrosine-based inhibitory and activating motifs (ITIM and ITAM) respectively on their intracellular domains. These have contrasting roles in immunomodulation, whereby, the ITIM can supress the local immune response of a cell and ITAM can activate it (127). CEACAM3 is found exclusively on neutrophils and upon receptor binding, the ITAM can induce phagocytosis (128). This would normally be disadvantageous to any pathogen that could activate this receptor.

In addition to CEACAMs, PSGs form the sister branch in the CEA family. All these proteins lack a membrane anchor and are expressed in soluble form from trophoblasts. As the name suggests, these proteins are expressed in pregnancy, though very little is known regarding

their specific functions within the host, however are essential for successful pregnancy (129).

Of the many proteins within the CEA family only four have been identified as pathogen receptors (CEACAM1, 3, 5 and 6) and each of these uses their N-terminal domain as the pathogen-binding domain. The N-terminal domain is structurally very similar across the whole family of CEACAMs, therefore, the pathogens utilising these receptors as adhesion sites have evolved to do so in a very specific way. In addition, analogous CEACAMs from other species, such as murine do not appear to allow binding of the human pathogen receptors – this would allow pathogens to target humans in a very specific manner. The difference in the head domains between species is more diverse than the intrinsic difference across the CEACAMs from an individual species.

Compared to membrane bound CEACAMs, little is known regrading secretory CEACAM splice variants; however, secretory CEACAM1 has been indicated in tissue immunomodulation and angiogenesis (130, 131). Additionally, secreted CEACAMs could be a diversion tactic as a defence against pathogens, preventing cellular adhesion.

Table 1.2 | **The CEA family of receptors.**

Shown are the members of the known human CEA family with the number of splice variants for each and whether they have been observed as a pathogen receptor. ITAM – immunoreceptor tyrosine kinase activating motif. ITIM – immunoreceptor tyrosine kinase inhibitory motif. GPI – Glycosylphosphatidylinositol. [1] By number of mRNA transcript variants on NCBI. [2] As identified by current literature.

CEA family member	Splice variants[1]	Pathogen receptor[2]	Additional features
CEACAM1	12	+	Intracellular ITIM
CEACAM3	3	+	Intracellular ITAM-like
CEACAM4	1	-	Intracellular ITAM-like
CEA (CEACAM5)	3	+	GPI anchor
CEACAM6	1	+	GPI anchor
CEACAM7	2	-	GPI anchor
CEACAM8	1	-	GPI anchor
CEACAM16	1	-	No membrane anchor; two IgV-like domains
CEACAM18	1	-	
CEACAM19	74	-	Intracellular ITAM
CEACAM20	4	-	Truncated N-domain; intracellular ITAM
CEACAM21	4	-	
PSG1	5	-	No membrane anchor
PSG2	1	-	No membrane anchor
PSG3	1	-	No membrane anchor
PSG4	4	-	No membrane anchor
PSG5	1	-	No membrane anchor
PSG6 (PSG10)	2	-	No membrane anchor
PSG7	1	-	No membrane anchor
PSG8	1	-	No membrane anchor
PSG9	3	-	No membrane anchor
PSG11	1	-	No membrane anchor

1.3.2 Structural Properties

With the exception of CEACAM20, all CEA family members begin with one IgV-like domain followed either directly by a membrane anchor or an arbitrary number of IgC-like domains. The IgV-like domain (immunoglobulin 'variable' fold) is slightly larger than the IgC (immunoglobulin 'constant' fold) domains containing usually 9 β-sheet motifs compared to 7 in IgC. These β-sheets are labelled for easy identification: A, B, C, C', C'', D, E, F and G, where C' and C'' are only found in IgV domains (132). The positions of the β-strands and the structure of the CEACAM1 IgV-like domain is shown in **FIGURE 1.9**.

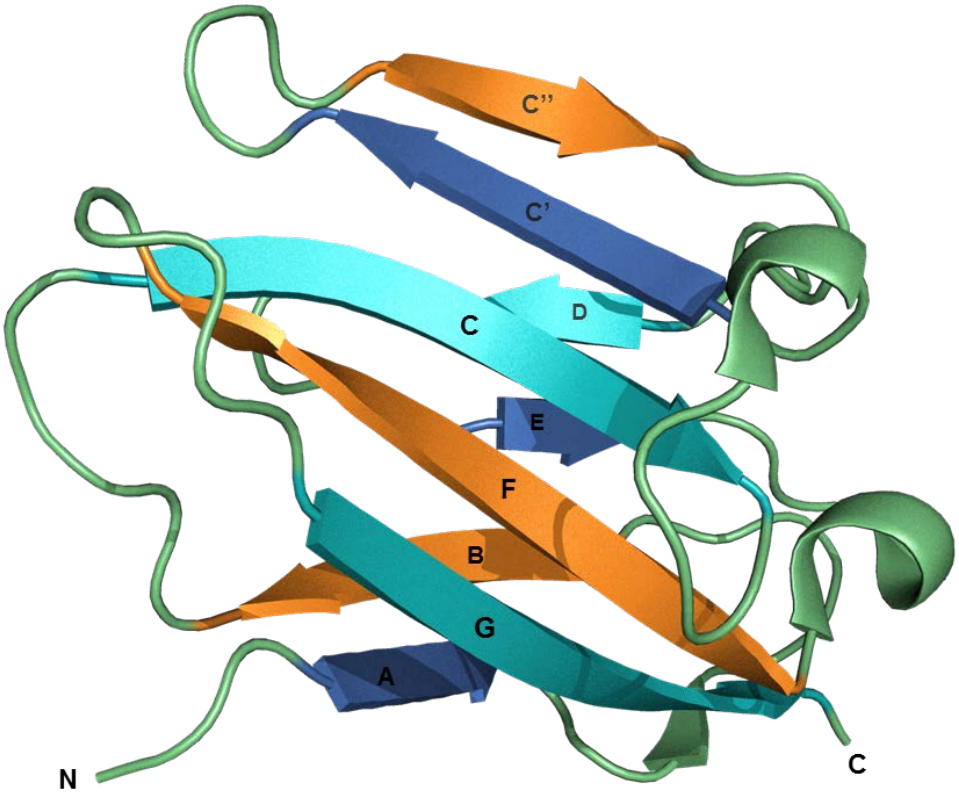


Figure 1.9 | **Structure of the CEACAM1 IgV-like domain.**

The crystal structure of the N-terminal IgV-like domain of CEACAM1 (PDB ID: 4WHD) contains 9 β-sheets each labelled as shown from A to G. The IgC fold lacks the C' and C'' sheets.

The two cases of CEA-family members that do not conform to the extracellular domain topology are CEACAM20 and CEACAM16. CEACAM20 harbours a severely truncated IgV-like domain and CEACAM16 contains two IgV-like domains flanking two IgC-like domains at the N and C termini. Moreover, CEACAM16 is the only member of the CEACAM branch not to contain a membrane-anchored splice variant, although all members of PSG branch do not either.

The length of the IgC-like repeating domains varies greatly between CEA family members as well as intrinsic splice variants, with CEACAM3, 4 and 19 lacking any IgC-like domains and CEA containing 6 IgC-like regions (**FIGURE 1.10**). The splice variant nomenclature states the total number of Ig-like domains (both IgV and IgC), as well as any other features which may be present/absent such as the membrane anchor or internal signalling domains, for example CEACAM-4L stands for CEACAM1 with 4 Ig-like domains and is the long variant form which contains the intracellular ITIM domain.

The major splice variant of CEACAM1 (CEACAM-4L) consists of one IgV- and 3 IgC-like domains in the extracellular unit. These domains are labelled as follows: N, A1, B and A2 from N- to C-terminus. Each domain is glycosylated via N-linked glycosylation similarly to other Ig-like proteins. This feature is important as it often overlooked in the crystal structures of recombinant CEACAMs expressed within *E. coli*, though it is not thought have much impact on pathogen-binding (133). It also contains the membrane spanning region and the intracellular ITIM domain.

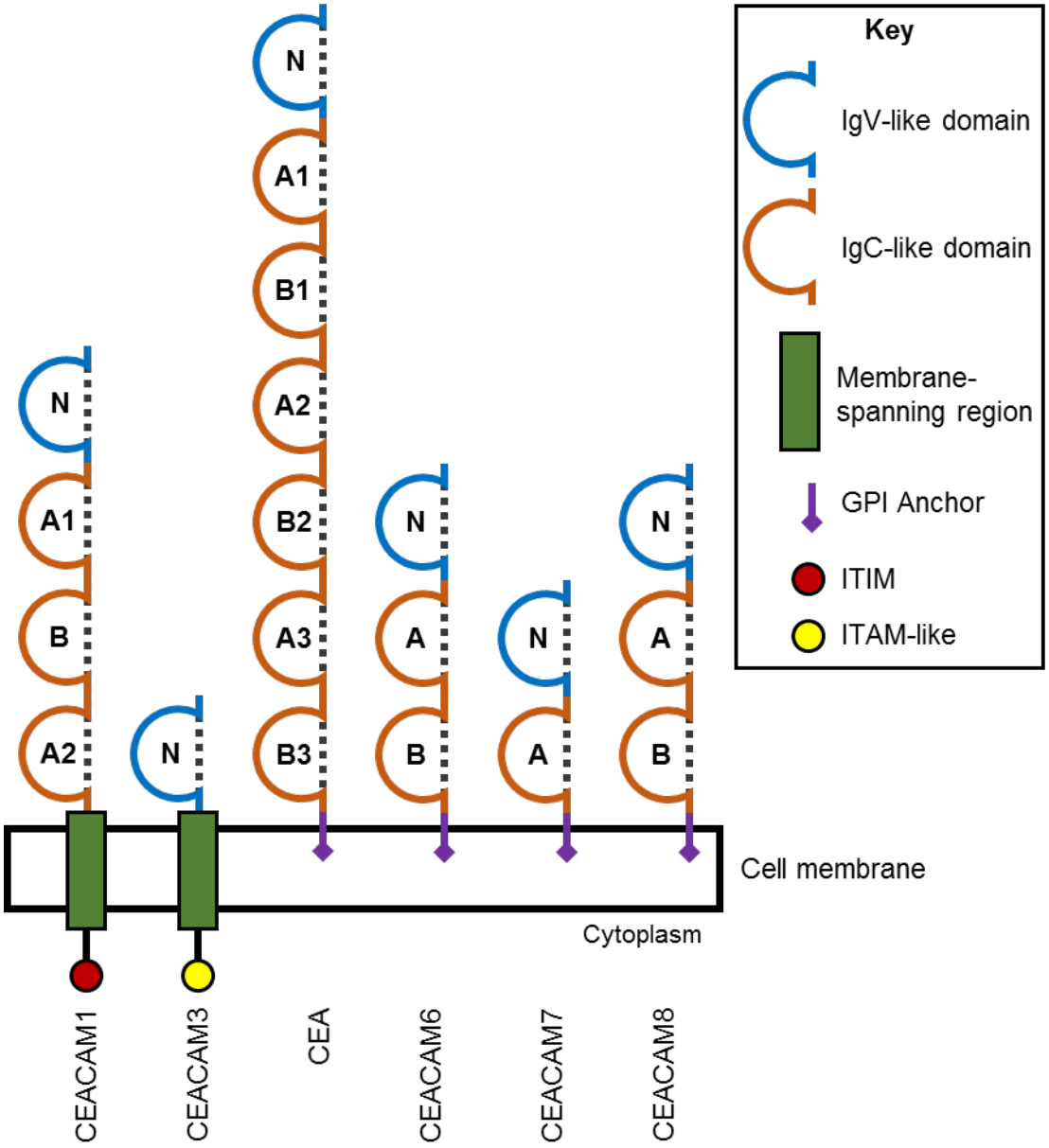


Figure 1.10 | **CEACAM topological domain diagram.**

The domain topology of the major splice variants of CEACAM1, CEACAM3, CEA, CEACAM6, CEACAM7 and CEACAM8 are displayed. GPI – glycophosphatidylinositol; ITIM – immunoreceptor tyrosine-based inhibition motif; ITAM – immunoreceptor tyrosine-based activation motif.

The predicted binding face of the N-terminal domain of CEACAM1 maps very closely to other CEACAMs, such as 3, 5 and 6, therefore, proteins such as UspA1 can bind to these proteins as well (118, 134). The pathogen binding surface utilises the C, C', C'', F and G β-sheet regions on the IgV-like fold and is given the name CFG face for brevity. Through computational and mutagenesis studies (133), key residues on the CFG face of CEACAM1 were identified that played a role in adhesion to UspA1: Y34, G47 and I91 found on the C, C' and F strands respectively. Other mutants were studied but were found to have a lesser to no impact on binding affinity (residue and domain location): S32 (C), V39 (loop between C and C'), Q44 (C') and V96 (loop between F and G) – though only alanine substitutions were examined for these residues. T56 (C'') and Q89 (F) were studied using more substitutions, but T56A, D and L showed little effect whereas Q89N appeared to increase binding affinity to UspA1 (133).

The Y34 residue is completely conserved across the entire CEA family and is likely involved in maintaining structural integrity – though it may increase affinity for binding proteins using hydrophobic interactions. CEACAM20 is an exception to this where it has a severely truncated N-terminal domain and lacks this residue entirely. However, CEACAM4, interestingly has a histidine at the equivalent position, making it the only member of the CEA-family to have this position and not have a tyrosine here.

In addition to the extracellular region, most members of the CEACAM branch have at least one splice variant that either possesses an alpha helical membrane-spanning region or a glycosylphosphatidylinositol (GPI) anchor. However, CEACAM16 and all PSGs do not possess any membrane anchor. As previously mentioned, some CEACAMs, such as 1 and 3, contain an intracellular signalling domain, but so does CEACAM4, CEACAM19 and CEACAM20. CEACAM1 is the only family member to possess an ITIM domain, where CEACAM3 and 4 contain an ITAM-like endocytic domain and CEACAM19 and 20 have a classical ITAM.

1.3.3 Pathogen Interplay and Disease

Since CEACAM1 has the potential to confer an immune inhibitory response when activated, it could therefore be advantageous for a pathogen to bind this receptor. Conversely, it would be to a pathogen's disadvantage to bind CEACAM3. Many bacteria do in fact bind to CEACAM1; *Neisseria meningitidis, Neisseria gonorrhoeae, Moraxella catarrhalis* and *Haemophilus influenzae* all harbour adhesins that specifically target human CEACAM1. The specificity for human CEACAM1 by these particular bacteria may explain why some of these pathogens cause disease exclusively in humans. The TAA UspA1 from *Moraxella catarrhalis* is known to bind to CEACAM1 facilitating bacterial adhesion. This protein is unique among known CEACAM-binding proteins as it utilises a coiled-coil stalk to adhere to CEACAM1 (114, 135, 136).

A separate study examined bacterial-induced signalling in the TIGIT receptor and its effect on the prevention of Natural Killer (NK) cells killing tumour cell lines (45). Like CEACAM1, this receptor contains an ITIM domain. In this case it was found that bacteria binding this receptor could activate the local immunosuppression response through ITIM in the tumour cells, preventing killing by NK cells.

As briefly mentioned, CEACAMs are primarily involved with cell-cell adhesion by interacting in a homo- or heterotypic fashion, for example, CEACAM1 will for homodimers though binding of the N-terminal domain and CEACAM6 and CEACAM8 can form heterodimers. CEACAM dimers also exist within the same membrane where the two distal N-domains can form dimers on the same cell and CEACAM1, for example, predominantly exits in homo-dimeric form (137). All known pathogen adhesins that bind CEACAMs utilise this same dimerization interface (CFG face), for example HopQ [*H. pylori*], Opa [*N. meningitidis*], UspA1 [*M. catarrhalis*] and Dr adhesins [*E. coli*] (114, 133, 138, 139).

The adhesion of pathogens to these receptors can have many different consequences, from passive adhesion to direct involvement in pathogenesis, for example, when the HopQ protein from *H. pylori* binds CEACAM1, this allows for the subsequent translocation of the

CagA oncoprotein into the target cell, which can lead to an increased risk of tumourigenesis (140). Other consequences can be immune evasion, as previously mentioned, via signalling using the ITIM domain on CEACAM1, such as with the Opa proteins from *N. meningitidis*.

There also exist certain genetic diseases with mutations in the CEACAM genes. A mutant variant of the CEACAM16 gene can lead to hearing disorders in later life as this protein is specifically expressed in the auditory system (141-144).

The GPI anchor of CEA (CEACAM5) is thought to have a role in cancer progression – it can inactivate the intrinsic cell-death pathway hence inhibiting anoikis (145) and CEA is a known biomarker for progression of colorectal cancer (146). However, the specific role of CEA in cancer remains vague.

1.4 *Fusobacterium* Vaccine Antigen Targets

Recently, TAAs have been used as successful vaccine antigens, such as the case for NadA [*N. meningitidis*] in a serogroup B meningococcal vaccine (147). This vaccine candidate was first proposed over 10 years prior to successful development of the full multicomponent vaccine. It was ideal as it was widely distributed and overexpressed in more virulent strains making it an ideal target (148).

Like NadA, the ubiquitous spread of TAAs throughout *Fusobacterium* represent potential future vaccine candidates. As the selection and sequences of TAAs is heavily based on the species, there would be little cross-reactivity between the harmful and more benign species, for example *F. necrophorum* compared to *F. periodonticum*. As mentioned in **Section 1.1.3.5**, there has been a number of attempts at cultivating a successful vaccine to prevent footrot, however the efficacy of these vaccines has been limited and the transition to equivalent human alternatives is doubtful. Therefore, a new approach at specifically targeting bacterial surface antigens may provide a promising avenue of research.

1.5 Aims

The primary aims of this study are to characterise the interactions between *Fusobacterium* and CEACAMs through the use of genomic, functional and structural analyses. The three aspects will combine to give an overall picture of the importance of these interactions and how it may vary from species to species.

Using whole-genome sequencing data, we will firstly clarify the species definition of *Fusobacterium*, which has been a topic of discussion recently (22, 23). The clarification of the species should provide an insight into the large disease variance seen between separate strains throughout the genus. In addition, we will sequence various uncharacterised clinical strains that were isolated from a range of diseases and examine their CEACAM-binding profiles and how this relates to *Fusobacterium* species.

In addition, we will characterise the proteins involved in CEACAM binding, labelled CbpFs, both from a functional and structural perspective using recombinant constructs. We will determine specificity to CEACAMs through binding assays and mutagenesis studies. This should help to identify features that may be important in disease progression and potentially provide useful information with respect to future vaccine design.

Chapter 2: Materials and Methods

2.1 Bacterial strains and growth conditions

Unless otherwise stated, *E. coli* strains were grown in Luria-Bertani (LB) broth at 37 °C with shaking at 200 RPM, or on LB agar plates supplemented with relevant antibiotics and compounds at the following concentrations: 100 $\mu g \cdot ml^{-1}$ ampicillin (Amp); 50 $\mu g \cdot ml^{-1}$ kanamycin (Kan); 34 $\mu g \cdot ml^{-1}$ chloramphenicol (Cm); 100 $\mu g \cdot ml^{-1}$ Zeocin™ (Zeo); 1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG); 20 $\mu g \cdot ml^{-1}$ 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-Gal). *Fusobacterium* spp. were grown in Fastidious Anaerobe broth (FAB) or on Fastidious Anaerobe agar (FAA) at 37 °C under anaerobic conditions: 90 % N_2, 5 % CO_2 and 5 % H_2 obtained in an anaerobic jar with a 2.5 L Oxoid™ AnaeroGen™ Sachet

(Thermo). Both static and shaking conditions were used for liquid cultures. Bacterial strains used in this work are listed in **TABLE 2.1**. Bacterial stocks were maintained at -80 °C in growth medium (LB broth or FAB) containing 25 % (v/v) glycerol.

Table 2.1 | **Stock bacterial strain list.**

The clinical strains stated use the source reference identifier as the strain name. [1] Made chemically competent. [2] Anaerobe Reference Unit (ARU), Public Health Wales Microbiology Cardiff, University Hospital of Wales, Cardiff CF14 4XW. [3] School of Oral and Dental Sciences, University of Bristol, Bristol BS1 2LY.

Strain Name	**Source**
Stellar™ *E. coli*	Clontech
BL21 (DE3) *E. coli* [1]	Sigma
BL21 (DE3) pLysS *E. coli*	Promega
Rosetta™ 2 (DE3) pLacI *E. coli*	Novogen
XL10-Gold® *E. coli*	Agilent
F. nucleatum ATCC 25586^T	ATCC
F. nucleatum ATCC 49256^T	ATCC
F. nucleatum ATCC 10953^T	ATCC
F. nucleatum R32310	ARU [2]
F. nucleatum R31249	ARU
F. nucleatum R30927	ARU
F. nucleatum R30464	ARU
F. nucleatum R32935	ARU
F. nucleatum R29976	ARU
F. nucleatum R26872	ARU
F. nucleatum R24394	ARU
F. nucleatum R18932	ARU
F. nucleatum R33458	ARU
F. nucleatum R30604	ARU
F. nucleatum R28385	ARU

F. nucleatum R28211	ARU
F. nucleatum R33533	ARU
F. nucleatum R18528	ARU
F. nucleatum R16531	ARU
F. nucleatum R5001	ARU
F. nucleatum R28427	ARU
F. nucleatum R28400	ARU
F. nucleatum R15792	ARU
F. nucleatum 2B3	SODS [3]
F. nucleatum 2B2	SODS
F. nucleatum 2B16	SODS
F. nucleatum 2B17	SODS
F. nucleatum 2B4	SODS

2.2 Eukaryotic strains and growth conditions

HeLa cell lines used were grown in Roswell Park Memorial Institute medium 1640 (RPMI-1640; Sigma) containing the following antibiotics and supplements: 100 $\mu g \cdot ml^{-1}$ penicillin (Pen); 100 $\mu g \cdot ml^{-1}$ streptomycin (Strep); 300 $\mu g \cdot ml^{-1}$ L-glutamine; and 10 % (v/v) foetal bovine serum (FBS). COS-1 cells were grown in Dulbecco's Modified Eagle Medium (DMEM; 1 $g \cdot l^{-1}$ glucose; Sigma) containing identical quantities of antibiotics and supplements as with RPMI. FBS is replaced with Gibco® Insulin-Transferrin-Selenium (ITS-G; Thermo) media supplement for low-serum (< 5 % [v/v] FBS) and serum-free requirements. CHO cells were grown in Ham′s Nutrient Mixture F12 (Sigma) with the addition of the same supplements though lacking serum. All cells were grown at 37 °C with 5 % CO_2 and were passaged before cells became confluent using a tryptic digestion.

Table 2.2 | **Eukaryotic cell lines used.**

Strain	Growth Media
$HeLa_{Neo}$	RPMI 1640
$HeLa_{CEACAM1}$	RPMI 1640
$HeLa_{CEACAM3}$	RPMI 1640
$HeLa_{CEA}$	RPMI 1640
$HeLa_{CEACAM8}$	RPMI 1640
COS-1	DMEM
CHO	Ham's F12

2.3 Gene Cloning

2.3.1 Creating a Plasmid for Producing Soluble Recombinant Protein

Two expression modes were used in assessing CbpF-CEACAM binding – soluble extracellular protein and surface expressed protein.

To produce recombinant soluble CbpF proteins, the genes encoding residues for extracellular protein, minus signal peptide, were cloned into pOPINE using ligation-independent cloning (LIC; In-Fusion® kit; Clontech). The pPOPINE plasmid encodes a hexa-histidine tag at the translated protein C-terminus (pOPINE was a gift from Ray Owens (149); Addgene plasmid 26043; **FIGURE S 1**). The plasmid was linearized with the restriction enzymes NcoI-HF® (New England Biolabs [NEB]) and PmeI (NEB) at 37 °C for 1 hr and the linearized plasmid DNA was purified with a QIAquick® PCR Purification Kit (Qiagen) according to the manufacturer's instructions to exchange buffers and remove the restriction enzymes.

The CbpF region of interest was amplified using the polymerase chain reaction (PCR) with primers listed in **TABLE 2.4** that had 5' DNA regions which were homologous with the ends of the linear plasmid for use in LIC. PCR was performed using CloneAmp™ high fidelity PCR premix (containing deoxynucleoside triphosphates (dNTPs), enzyme and additives;

Clontech), 0.5 µM primers and 1 ng·µl^{-1} genomic DNA. The conditions used in PCR were as follows: initial 95 °C 10 min, 95 °C 10 sec, 60 °C 10 sec (some primers required for higher/lower temperatures to anneal), 72 °C 5 sec·kbase^{-1} DNA, repeat from step 2 30 times, lastly 72 °C 10 min for the final extension.

The linearized plasmid and PCR were incubated at 50 °C for 15 min with 1X In-Fusion® master mix at a 2:1 molar ratio of PCR to plasmid. The In-Fusion® reaction was then incubated with chemically competent Stellar® *E. coli* cells (Clontech) for 30 min on ice. The cells were subsequently heat-shocked by incubating them at 42 °C for 40 secs followed by 5 min on ice, before the addition of preheated (37 °C) SOC medium (Super Optimal broth with Catabolite repression) at a ratio of 1:10 cell suspension to SOC. The cells were then incubated with shaking at 37 °C at 200 RPM for 1 hr before plating onto LB agar plates (containing Amp, IPTG and X-Gal) and incubated overnight (O/N) at 37 °C.

Individual white colonies were picked and grown in 10 ml LB broth (containing Amp) O/N at 37 °C with shaking at 200 RPM before extracting the plasmid DNA using a QIAprep® Spin kit (Qiagen). The plasmid was adjusted to a concentration of 100 ng·µl^{-1} (as measured using a DeNovix® DS-11 FX UV-Vis spectrophotometer) and was sequenced using Sanger sequencing (Source BioScience, Nottingham) using the T7F and T7R standard primers. Sequences were aligned with the expected sequence and plasmids shown to have the correct gene of interest (GOI) were then transformed into expression cells: *E. coli* BL21 (DE3), *E. coli* BL21 (DE3) pLysS (Promega) and Rosetta™ 2 (DE3) pLacI (Novagen) using the heat shock method as previously described. The plasmids produced using pOPINE were labelled pCFR1 and pCBR1 for CbpFa and CbpFb respectively (**TABLE 2.3**).

In addition to using pOPINE in protein production, the pMAL-c5X plasmid (NEB; **FIGURE S 3**) was also used for producing fragments of CbpFa. This plasmid encodes maltose-binding protein (MBP) as well as a ploy-asparagine linker directly upstream of the GOI start codon. This plasmid was linearized using XmnI and BamHI-HF® (NEB) and purified with identical conditions to pOPINE. Constructs using this vector as a backbone are detailed in **TABLE 2.3**.

The genes inserted into this vector were amplified using the primers listed in **TABLE 2.4** which were designed for LIC using In-Fusion®. A hexa-histidine tag was also encoded within the 3' reverse primer, therefore providing an additional method of purification in downstream applications.

2.3.2 Creating a Plasmid for Surface Expression of Protein

The same cloning method was used as described in the previous section, however the plasmid pOAF was used. This plasmid encodes the *E. coli* OmpA signal peptide directly upstream of the pOPINF 5' addon sequence, which contains a hexa-histidine tag and a linker sequence prior to the GOI start codon. This plasmid was linearized using KpnI-HF® and HindIII-HF® (NEB) for 1 hr at 37 °C and subsequently purified using the QIAquick® PCR purification kit as previously explained.

The regions encoding CbpFa and CbpFb lacking the signal peptide were then amplified by PCR (using method described in **SECTION 2.3.1**) and cloned into pOAF using In-Fusion® LIC. Other TAAs from *Fusobacterium* were cloned into this vector as well as truncation mutants of CbpFa. All the plasmids created using pOAF and the primers used in PCR are listed in **TABLE 2.3** and **TABLE 2.4** respectively. After successful sequencing, pOAF-based plasmids were cloned into *E. coli* BL21 (DE3) pLysS (Promega) using the heat-shock method ready for expression.

2.3.3 Site-directed Mutagenesis

To create mutants of the CEACAM1 N-terminal IgV-like domain, a variant of site-directed mutagenesis was performed utilising the LIC with the InFusion® kit. Parent plasmid was transformed into XL10-Gold® *E. coli* competent cells (Agilent), prior to extraction and purification using a QIAprep® Spin kit. These cells were used instead of Stellar™ cells as they methylate their DNA. Primers were designed where the forward primer contained to desired mutation as well as 15 bp of complementary DNA (on the 5' end) to the 5' end of the reverse primer. The reverse primer was directly upstream of the forward primer, such that the whole plasmid would be amplified. After plasmid amplification using PCR and

verification on an agarose gel, the reaction was purified using a QIAquick® PCR purification kit followed by parent plasmid digestion using DpnI (NEB) according to the manufacturer's protocol. This reaction was purified using the PCR purification kit again and 1 µl 5X InFusion® master mix was added to 1 µl of digested PCR and 3 µl of water. The reaction was incubated for 15 min at 50 °C before transforming into Stellar™ cells using the heat shock method. Colonies were selected the following day and 10 ml liquid cultures were set up O/N at 37 °C and the plasmids purified using a miniprep kit (Qiagen). Purified plasmids were then sent for sequencing to confirm the presence of the desired mutations. Primers for these constructs are detailed in **TABLE 2.4**.

2.3.4 Plasmids and Primers

Table 2.3 | **Plasmids List**.

A list of all plasmids used and created in this study. Plasmid maps are detailed in **APPENDIX C**. hCC1 – human CEACAM1. [1] pINFUSE-hIgG1-F_C2 is the full name. [2] The start residue is 22 and the end residue number is 330. [3] **APPENDIX B** clarifies the CEACAM1 numbering convention.

Vector Name	Backbone	CDS Product	Resistance	Source
Stock Plasmids				
pOPINE	pTriEx-2		Amp^R	(149)
pOAF	pOPINF	OmpA signal peptide	Amp^R	*Inhouse*
pcDNA3.1-CC1	pcDNA3.1(+)	hCC1 NA1BA2-F_C	Amp^R, Neo^R	*Inhouse*
pMAL-c5X		MBP	Amp^R	NEB
pINFUSE2[1]		IL2 signal peptide	Zeo^R	Invivogen
Soluble Protein Expression Plasmids				
pCFR1	pOPINE	CbpFa 22-330	Amp^R	*This study*
pCBR1	pOPINE	CbpFb 25-374	Amp^R	*This study*
pCFR2	pOPINE	CbpFa 22-128	Amp^R	*This study*
pCFR3	pOPINE	CbpFa 128-180	Amp^R	*This study*
pCFR4	pOPINE	CbpFa 180-235	Amp^R	*This study*
pCFR5	pOPINE	CbpFa 214-330	Amp^R	*This study*
pCFR6	pOPINE	CbpFa Δ148-179[2]	Amp^R	*This study*
pCFR7	pOPINE	CbpFa 22-283	Amp^R	*This study*
pCFM1	pMAL-c5X	MBP-CbpFa 40-331	Amp^R	*This study*
pCFM2	pMAL-c5X	MBP-CbpFa 40-190	Amp^R	*This study*
pCFM3	pMAL-c5X	MBP-CbpFa 120-240	Amp^R	*This study*
pCFM4	pMAL-c5X	MBP-CbpFa 180-331	Amp^R	*This study*
Surface Expression Plasmids				
pCFS1	pOAF	CbpFa 22-479	Amp^R	*This study*
pCBS1	pOAF	CbpFb 25-519	Amp^R	*This study*
pTAAS1	pOAF	FN0471 17-240	Amp^R	*This study*
pTAAS2	pOAF	FN0735 25-602	Amp^R	*This study*

pTAAS3	pOAF	FNP1391 25-644	Amp^R	*This study*
pCFS2	pOAF	CbpFa 111-479	Amp^R	*This study*
pCFS3	pOAF	CbpFa 180-479	Amp^R	*This study*
pCFS4	pOAF	CbpFa 214-479	Amp^R	*This study*
pCFS5	pOAF	CbpFa 293-479	Amp^R	*This study*
pCFS6	pOAF	CbpFa 329-479	Amp^R	*This study*
Mammalian Protein Expression Shuttle Vectors				
pCCR3	pINFUSE2	hCC1 NA1B-F_C	Zeo^R	*This study*
pCN29G	pCCR3	hCC1 NA1B F29G-F_C[3]	Zeo^R	*This study*
pCN29I	pCCR3	hCC1 NA1B F29I-F_C	Zeo^R	*This study*
pCN29R	pCCR3	hCC1 NA1B F29R-F_C	Zeo^R	*This study*
pCN29Y	pCCR3	hCC1 NA1B F29Y-F_C	Zeo^R	*This study*
pCN44E	pCCR3	hCC1 NA1B Q44E-F_C	Zeo^R	*This study*
pCN44L	pCCR3	hCC1 NA1B Q44L-F_C	Zeo^R	*This study*
pCN44R	pCCR3	hCC1 NA1B Q44L-F_C	Zeo^R	*This study*

Table 2.4 | **Primers used to create plasmids.**

Both forward and reverse primers are listed for all the constructs made in this study. Regions of homology required for LIC are underlined and primer mismatches encoding specific mutations are highlighted in bold. [1] gDNA – genomic DNA extracted and purified from parent strain. [2] Two primer pairs are listed for this construct as it requires two PCRs, digestion and a ligation to create the insert for the final plasmid. The XhoI restriction site is shown in italics. [3] Reverse primers for the pMAL-c5X constructs also encode a hexa-histidine tag.

Primer Sets (5' – 3')	**Used to Create**	**Template**
AGGAGATATACCATGTCTTATTCAGCTGCACCAGTTATT GTGATGGTGATGTTTACCAGTGCCAAGCTTAGCTT	pCFR1	*Fn* ATCC 25586 gDNA[1]
AGGAGATATACCATGGCCCCAGCATTTGGAACA GTGATGGTGATGTTTAGCAGAACCTCCCCCTGT	pCBR1	2B3 gDNA
AGGAGATATACCATGTCTTATTCAGCTGCACCAGTTATT GTGATGGTGATGTTTATATTGACTTCCAAAAGCTGAACTATTA	pCFR2	*Fn* ATCC 25586 gDNA
AGGAGATATACCATGTATCAAGTTACTGGAAACTTTTCT GTGATGGTGATGTTTATGAACATTATTACCTAAAATAAAGTTATC	pCFR3	*Fn* ATCC 25586 gDNA
AGGAGATATACCATGCATATTGGCGGTGGTATTAATAATTC GTGATGGTGATGTTTTAATTGTCTACCAGTAACAGCATC	pCFR4	*Fn* ATCC 25586 gDNA
AGGAGATATACCATGATAGTTAATGTTGGAGATGGAGCTAT GTGATGGTGATGTTTACCAGTGCCAAGCTTAGCTT	pCFR5	*Fn* ATCC 25586 gDNA
AGGAGATATACCATGTCTTATTCAGCTGCACCAGTTATT TAATAA*CTCGAG*ATACTGACCATTGAATTCACCCATT TAATAA*CTCGAG*CATATTGGCGGTGGTATTAATAATTCAGTAG GTGATGGTGATGTTTACCAGTGCCAAGCTTAGCTT	pCFR6[2]	*Fn* ATCC 25586 gDNA
AGGAGATATACCATGTCTTATTCAGCTGCACCAGTTATT GTGATGGTGATGTTTAGGAGCTCCTCCACCTCCACCAGAGC	pCFR7	*Fn* ATCC 25586 gDNA
AAGTTCTGTTTCAGGGCCCGTCTTATTCAGCTGCACCAGTTATTA ATGGTCTAGAAAGCTTTATTTATTTTTTAATAACATATTTAAC	pCFS1	*Fn* ATCC 25586 gDNA
AAGTTCTGTTTCAGGGCCCGGCCCCAGCATTTGGAACAGGAACAG ATGGTCTAGAAAGCTTTATTTATTTTTTAAATTCATATTTAAT	pCBS1	2B3 gDNA

AAGTTCTGTTTCAGGGCCCGGCACCAACTATTGAGGCAG ATGGTCTAGAAAGCTTTATTTAGTTTTCAATAATTTATTTAACTTTTC	pTAAS1	*Fn* ATCC 25586 gDNA
AAGTTCTGTTTCAGGGCCCGGCTACACCAACTATTGAAGC ATGGTCTAGAAAGCTTTATTTATTTTTTAATAGTCTATTTAATTTTTCT	pTAAS2	*Fn* ATCC 25586 gDNA
AAGTTCTGTTTCAGGGCCCGGCTACTCCAACTATTGAAG ATGGTCTAGAAAGCTTTATTTATTTTTTAATAGTCTATTTAATTTTTC	pTAAS3	*Fnp* ATCC 10953 gDNA
TCGGGATCGAGGGAAGGGCAGGAGTTGATAATGTAG CCTGCAGGGAATTCGGATCTTAGTGATGGTGATGGTGATGTTTAG CACCAGTGCCAAGCTT	pCFM1[3]	*Fn* ATCC 25586 gDNA
TCGGGATCGAGGGAAGGAATAGTTCAGCTTTTGGAAGT CCTGCAGGGAATTCGGATCTTAGTGATGGTGATGGTGATGTTTAG CTACTGAATTATTAATACCAC	pCFM2	*Fn* ATCC 25586 gDNA
TCGGGATCGAGGGAAGGCATATTGGCGGTGGTATTAA CCTGCAGGGAATTCGGATCTTAGTGATGGTGATGGTGATGTTTTCC ATTTCCACTATATAATTGTCTA	pCFM3	*Fn* ATCC 25586 gDNA
TCGGGATCGAGGGAAGGCATATTGGCGGTGGTATTAA CCTGCAGGGAATTCGGATCTTAGTGATGGTGATGGTGATGTTTAG CACCAGTGCCAAGCTT	pCFM4	*Fn* ATCC 25586 gDNA
CTTGTCACGAATTCGATAGGGCACCTCTCAGCCCCA GTGAGTTTTGTCAGATCTAGTGACTATGATCGTCTTGAC	pCCR3	pcDNA3.1-CC1
CAATCTGCCCCAGCAACTT**GG**TGGCTACAGCTGG AAGTTGCTGGGGCAGATTGTGGACAAGGAG	pCN29G	pCCR3
CAATCTGCCCCAGCAACTT**A**TTGGCTACAGCTGG AAGTTGCTGGGGCAGATTGTGGACAAGGAG	pCN29I	pCCR3
CAATCTGCCCCAGCAACTT**CG**TGGCTACAGCTGG AAGTTGCTGGGGCAGATTGTGGACAAGGAG	pCN29R	pCCR3
CAATCTGCCCCAGCAACTTT**A**TGGCTACAGCTGG AAGTTGCTGGGGCAGATTGTGGACAAGGAG	pCN29Y	pCCR3
GAGTGGATGGCAACCGT**G**AAATTGTAGGATATGC ACGGTTGCCATCCACTCTTTCCCCTTTG	pCN44E	pCCR3
GAGTGGATGGCAACCGTC**T**AATTGTAGGATATGC ACGGTTGCCATCCACTCTTTCCCCTTTG	pCN44L	pCCR3

GAGTGGATGGCAACCGTC**G**AATTGTAGGATATGC ACGGTTGCCATCCACTCTTTCCCCTTTG	pCN44R	pCCR3

2.4 Bacterial Protein Expression and Purification

Expression and purification methods were optimised on a case-by-case basis, which will be explained where necessary. The most common and reliable methods used are explained here in detail and primarily pertain to the expression of the full-length constructs of CbpFa and CbpFb.

2.4.1 Small-scale Expression and Native Purification

Transformed *E. coli* BL21 (DE3) pLysS cells were grown in LB (< 1 l) containing Amp and Cm until mid-log phase, $OD_{600nm} = 0.4$, where they were induced with 1 mM IPTG and grown for a further 3 hours at 37°C. Cells were harvested by centrifugation at 6000 x *g* for 10 min and were resuspended in Lysis Buffer (Native Buffer A [**TABLE S 1**] with 1X protease inhibitor cocktail [**TABLE S 1**], 1 U·ml^{-1} DNAse I [Thermo], 2 mM $MgCl_2$ and 100 µg·ml^{-1} lysozyme [Sigma]). The volume (ml) of lysis buffer used was 5 times the mass (g) of the cell pellet. The suspension was incubated on ice for 30 min before sonicating. Lysates were spun at >10000 x *g* for 1 hr and the supernatant was retained.

His-tagged protein was purified using Ni-NTA agarose (Qiagen). The cell lysate supernatant was incubated with Ni-NTA agarose (1:50 ratio agarose to lysate) for 1 hr at 4 °C with gentle mixing. The agarose was pelleted by centrifuging the tube for 5 min at 200 x *g*. The supernatant was carefully decanted and retained for downstream analysis. The remaining slurry was loaded onto a 5 ml polypropylene gravity flow column. 20 column volumes (CV) of Native Buffer A (**TABLE S 1**) were flowed through the column under gravity. 10 CV of the same buffer containing 150 mM imidazole was then run down the column to wash off non-specifically bound proteins. 5 x 1 CV size elution fractions were collected using Native Buffer B (**TABLE S 1**).

Fractions were run on an 4-20% acrylamide gel and stained with Coomassie Quick Stain (Generon) and assessed for purity. Pure protein fractions were then exhaustively dialysed (> 100:1 dialysate to fraction; 3 buffer changes with > 4 hours between each, and one O/N at 4°C) against the dialysate of choice, e.g. 20 mM Trizma®-HCl pH 7.5, 100 mM NaCl, 1 % (v/v) glycerol or phosphate buffered saline (PBS; **TABLE S 1**). Protein was then concentrated to a sufficient concentration using Vivaspin® 20 spin columns (Sartorius; protein concentrations were measured using a DeNovix® DS-11 FX UV-Vis spectrophotometer with predicted Mr and Ext. coefficient).

2.4.2 Large-scale Expression and Native Purification

Rosetta™ 2 (DE3) pLacI cells containing pOPINE with the GOI were inoculated into 10 ml LB broth containing Amp and Cm and grown O/N at 37 °C with shaking at 200 RPM. This starter culture was then added to total of 8 l autoinduction terrific broth (Formedium), containing Amp and Cm, before incubating for 24 hrs at 37 °C with shaking 200 RPM. Cells were pelleted and resuspended in Lysis Buffer and sonicated as described previously. The cell lysate was applied to a 5 ml HisTrap FF (GE Healthcare) column charged with NiCl as according to the manufacture's procedures and equilibrated with Native Buffer A (**TABLE S 1**). The flow through was retained and the column was then connected to an ÄKTA and washed with Native Buffer A with a flow rate of 1 ml·min^{-1} until a stable $UV_{280\,nm}$ absorbance was attained. A gradient was then setup with a target of 100 % Native Buffer B (**TABLE S 1**) over a time course of 1 hour.

Fractions where a clear $UV_{280\ nm}$ absorbance peak was observed were retained and analysed on a 4-20% SDS-PAGE gel. Fractions with protein bands at the correct size were pooled, concentrated and dialysed against SEC Buffer A (**TABLE S 1**) in preparation for size-exclusion chromatography (SEC). A gel filtration column (ProteoSEC 3-70 kDa 16x600 mm; Generon) was equilibrated with SEC Buffer A before loading the pooled sample and running with a flow rate of 0.5 ml·min^{-1}. Fractions with a strong $UV_{280\ nm}$ absorbance peak were

analysed on a 4-20% SDS-PAGE gel, pooled and adjusted to the desired concentration for use in downstream applications.

2.4.3 Large-scale Expression and Denaturing Purification

For purifying protein under denaturing conditions, the protocol remained largely the same as under native conditions, however, the cell resuspension and lysis is achieved using Denaturing Buffer A (**TABLE S 1**) which contains 8 M urea with an incubation at room temperature (RT) for 1 hr with gentle agitation. The lysed and denatured solution was then spun at 10000 x *g* for 1 hr and the supernatant retained. The supernatant was passed through a HisTrap FF column charged with Ni^{2+} and protein purified as previously described, though replacing Native Buffer B with of Denaturing Buffer B (**TABLE S 1**). Protein containing fractions were pooled and concentrated. During concentration using a spin column, the urea concentration was reduced in a stepwise manner using concentrations of 6, 4, 2 and 1 M urea and the resulting solution was centrifuged at >10000 x *g* for 20 min and the supernatant was exhaustively dialysed against SEC Buffer A (**TABLE S 1**). The solution was spun for a further 5 min at 10000 x *g* to remove precipitants prior to loading onto the gel filtration column where the procedure is identical to native conditions.

2.4.4 Surface Expression

To produce surface expressed proteins, cells harbouring the pOAF-based plasmids were grown overnight at 37 °C before being diluted in fresh LB broth and grown at 37 °C O/N in the presence of 1 mM IPTG, after reaching $OD_{600\,nm} = 0.4$, prior to use in binding assay experiments.

2.5 *Fusobacterium* Lysate Preparation

Fusobacterium strains were grown for 2 days under anaerobic conditions as described in **SECTION 2.1**, before being resuspended in PBS (**TABLE S 1**) containing protease inhibitors (**TABLE S 1**) and adjusting to an $OD_{280\,nm} = 2$. The cell suspensions were then freeze-thawed

three times to lyse the cells and SDS-PAGE loading buffer (**TABLE S 1**) was added to a final concentration of 1X. Lysates were stored at -20 °C.

2.6 Human CEACAM IgG1-F_C fusion protein production

2.6.1 DNA Preparation

For small-scale transfections, plasmid DNA was purified from *E. coli* XL10-Gold® cells using a miniprep kit (Qiagen) to yield >20 µg DNA from a 10 ml O/N culture. For large-scale transfections, 1 l cultures of *E. coli* were setup and grown to O/N before preparing the plasmid DNA using an EndoFree® Plasmid Maxi kit (Qiagen) to yield up to 10 mg DNA.

2.6.2 Large Scale Transient Transfection

COS-1 cells were grown in T175 tissue culture flasks to 70-90% confluency. The cells were washed with copious amounts of Dulbecco's phosphate buffered saline (DPBS) to remove excess serum. The cells were then transfected using Lipofectamine® 2000 (Thermo) according to the manufacture's protocols using serum-free media throughout. Supernatants were taken at days 3 and 7 post transfection and analysed by dot-blotting using anti-human F_C to detect the presence of the recombinant peptide. The resulting positive supernatants were pooled, filtered through a 0.2 µm syringe filter (Sartorius) and concentrated to 10 ml using centrifugal concentrators (14 ml capacity; 5000 kDa MWCO; Sartorius).

Concentrated supernatants were diluted 1:1 with Protein A Loading Buffer (**TABLE S 1**) and loaded onto a column containing 1 ml Protein A-Sepharose® resin (Sigma; pre-equilibrated with loading buffer) and allowed to run through under gravity. The column was washed with 20 ml loading buffer to thoroughly remove excess unbound proteins. Protein was eluted into 1 ml fractions using Protein A Elution Buffer (**TABLE S 1**; 200 mM Na_2HPO_4, 100 mM citric acid, pH 3.0). The fractions were immediately neutralised with Protein A Neutralisation Buffer (**TABLE S 1** ; 1 ml 1 M Trizma®-HCl pH 7.5) and samples were collected and analysed using a Western blot using a CEACAM-specific primary antibody, AO115. Positive fractions

were then pooled and concentrated to a suitable volume and resulting protein concentration was quantified using a BCA assay (Thermo) according to the manufacture's procedures.

2.6.3 Small-scale Transfections and Purification

COS-1 cells were grown to 70-90 % confluency in 6-well tissue culture plates. DNA-lipid complexes were created using Lipofectamine 3000 (Thermo) and cells were transfected according to the manufacture's protocols. Supernatants were analysed for protein by immunodot blotting and positive wells were harvested at day 5, prior to purification.

Supernatants were diluted in a 1:1 ratio with Protein A Loading Buffer (**TABLE S 1**) and filtered through a 0.2 µm syringe filter. 100 µl Protein A Sepharose® was added per 5 ml diluted sample and incubated at RT for 1 hr. Samples were subsequently centrifuged for 5 min at 200 x *g*. The supernatant was removed (but retained for further analysis) leaving approximately 1 ml liquid to resuspend the resin pellet in. This was then loaded onto Pierce™ spin cups (Thermo; paper filters) and spun at 1000 x *g* for 1 min. All column flow-throughs were retained for analysis post-purification. 10 CV of Protein A Loading Buffer was passed through the columns by iterative centrifugation at 1000 x *g* for 1 min. 5 fractions with 1 CV Protein A Elution Buffer (**TABLE S 1**) was then passed through and immediately neutralised with 1:9 ratio of Protein A Neutralisation Buffer (**TABLE S 1**) to fraction volume. All fractions and flow-throughs were analysed for protein using immunodot blotting with an anti-human IgG1 F_C-AP antibody. Positive fractions were pooled and concentrated to < 100 µl.

As the protein yield from this method was very low (not measurable with BCA or NanoDrop) quantitative ELISA was used to determine the relative protein abundance as described in **SECTION 2.7.1**, often in the order of fM.

2.7 Western blots, Immunodot blots and ELISAs

Proteins were transferred to a nitrocellulose membrane either by direct pipetting and vacuum drying (immunodot blots), or by electrophoretic transfer from an acrylamide gel at

0.3 A for 1 hour (Western blots). For ELISAs, proteins are diluted in carbonate buffer (50 mM Na_2CO_3, 50 mM $NaHCO_3$ pH 9.6) to a suitable concentration (0.1 – 1 $\mu g \cdot ml^{-1}$) prior to incubation on an ELISA plate for 1 hour at RT or O/N at 4°C. The non-specific binding sites on the membrane/plate were then blocked with 3% (w/v) bovine serum albumin (BSA; diluted in PBS-T: PBS containing 0.05% [v/v] TWEEN®-20 [Sigma]). Primary antibodies are diluted from stock solutions (0.1-1 $\mu g \cdot ml^{-1}$) in 1% (w/v) BSA in PBS-T with 0.05% (w/v) NaN_3 and added to the membrane/wells at RT for 1 hour. The membrane/wells were then washed 3 times with PBS-T with the final wash incubated for 5 min before decanting. The secondary antibody formulation, addition and wash steps are identical.

The blots were developed using NBT/BCIP in AP buffer. NBT (4-nitro blue tetrazolium chloride) stock: 50 $mg \cdot ml^{-1}$ in 100% DMF (dimethylformamide); BCIP (5-bromo-4-chloro-3-indolyl-phosphate) stock: 50 $mg \cdot ml^{-1}$ in 70% DMF; AP (Alkaline phosphatase) Buffer (**TABLE S 1**). Developing buffer uses 66 µl NBT and 33 µl BCIP per 5 ml of AP buffer. The development was stopped by washing thoroughly with distilled water and the blots were then allowed to dry.

ELISAs were developed using SIGMAFAST™ p-Nitrophenyl phosphate tablets (Sigma) dissolved in distilled water and incubated at 37°C for sufficient time to see development (0.5-5 hours). The absorbance for each well was measured using a 96-well plate reader using a wavelength set to 405 nm.

2.7.1 Quantitative ELISA

To measure very low relative protein concentrations, quantitative ELISA (qELISA) was used. A protein of known concentration and equivalent immunogenicity (identical protein or close mutant) was added in range of concentrations from 100 fM to 100 nM (diluted in carbonate buffer). Unknown samples were diluted in a range of 1:10, 1:100 and 1:1000 sample to carbonate buffer. 100 µl for all conditions and the standards were added to wells in an ELISA plate in triplicate. The samples were then incubated O/N at 4 °C before

continuing the ELISA as previously described, using an identical specific primary antibody for the unknowns and standards.

After measuring absorbance at 405 nm, a standard curve was produced that adhered to a logarithmic curve, which formula was calculated using logarithmic regression. This was then used to calculate the unknown concentrations that were within the range of the standard curve.

2.8 Adhesion assays

2.8.1 Eukaryotic cell preparation

HeLa cells transfected with DNA conferring neomycin resistance and with the human CEACAM genes were grown to confluence at 37 ° in a T75 flask. The cells were trypsinised, collected, counted and diluted to attain a 20% confluence in a 96-well plate (approximately $2x10^4$ cells). 100 µl of cell suspension was added to each well and grown for 48 hours until confluent. The media was replaced with antibiotic free pre-warmed (37 °C) Medium 199 (Sigma) and washed a further 2 times to remove traces of antibiotics from the wells.

2.8.2 Inhibitor preparation

The recombinant peptides rD-7 and MsfA (kindly provided by Darryl Hill and Clio Andreae respectively) were diluted to a concentration of 10 $\mu g \cdot ml^{-1}$ in Medium 199 (at 37 °C) and 100 µl was added to the relevant wells. The plates were incubated for 30 minutes at 37 °C. The media was then removed prior to bacterial incubation.

2.8.3 Bacterial preparation

The bacteria to be tested were grown and induced as described earlier. The overnight cultures were diluted to $OD_{600nm} = 1$ in Medium 199 (at 37 °C) to reach an estimated multiplicity of infection (MOI) of 500 bacterial cells per 1 eukaryotic cell. All medium was removed from the wells and 100 µl of the bacterial suspension was then added. The plates were incubated for 3 hours at 37 °C.

2.8.4 Cell fixation

After incubation with bacteria, each well was washed thoroughly with pre-warmed Medium 199 four times to remove any unbound bacteria. The wells were then washed twice with warm DPBS. All liquid was removed from each well and 50 µl of 4% (w/v) paraformaldehyde (PFA; dissolved in PBS) was added. The plates were incubated at 4 °C O/N to fix. The PFA was removed and the wells washed twice with DPBS-Azide (DPBS containing 0.05% [w/v] Sodium Azide [NaN_3]) to remove any residual PFA to prevent interference with antibody staining.

2.8.5 Cell staining

The wells were blocked to prevent non-specific antibody binding to the well surface by adding 100 µl of 3% (w/v) BSA and incubating at RT for one hour with gentle agitation. The BSA was removed and 100 µl of rabbit anti-*E. coli* LPS (Biogenesis) diluted to 1 $\mu g \cdot ml^{-1}$ in 1% (w/v) BSA in PBS-T, containing 0.05% NaN_3, was added to each well and incubated for one hour at RT with gentle agitation. The primary antibody was aspirated off and the wells were washed thoroughly three times with PBS-T with the final wash step incubating in PBS-T for 5 minutes at RT with gentle agitation to remove residual antibody. 100 µl of the secondary antibody (Goat anti-Rabbit Alexa Fluor® polyclonal antibody conjugated to Fluorescein isothiocyanate [FITC; Thermo Scientific]) diluted to 1 $\mu g \cdot ml^{-1}$ in 1% (w/v) BSA in PBS-T with 0.05% NaN_3 was added to each well and incubated at RT with gentle agitation for one hour. The secondary antibody was removed, and the wells washed as before three times with PBS-T. 100 µl of 0.1% (v/v) Triton™ X-100 (diluted in distilled water) was added to each well and incubated at RT for 10 min to permeabilize the HeLa cells. The liquid was aspirated off and 100 µl of 4',6-diamidino-2-phenylindole (DAPI; Sigma) diluted to 1 $\mu g \cdot ml^{-1}$ in 1% (w/v) BSA in PBS-T with 0.05% NaN_3 was added and the plates incubated for 15 min at RT with gentle agitation. The wells were washed as before to remove any excess DAPI with PBS-T three times prior to the addition of 50 µl of PBS-Azide to each well.

2.8.6 Cell imaging

A photograph of representative cells within each well was taken using a fluorescent microscope (Olympus IX70) using wavelengths set at 360 nm and 500 nm to detect DAPI and FITC respectively. Bacteria fluoresced green and HeLa cell nuclei blue. Images were captured using a Hamamatsu C4247-95 ORCA 100 series camera and analysed using HCImage (release 4.3.1; Hamamatsu Corporation). Post analysis of images only examined the green channel and calculated total fluorescence.

2.9 Crystallography

Purified protein was concentrated to a concentration between 1-10 $mg \cdot ml^{-1}$, where the highest possible concentration was favoured prior to precipitation. Concentration was determined using a NanoDrop using predicted extinction coefficient and molecular weight, calculated using ExPasy ProtParam tool (150). Protein solutions were screened against a variety of high throughput crystallography screens with drop volume, ratio and protein concentration being varied as well. High throughput screens and proteins were dispensed, using a Crystal Phoenix robot (Art Robbins Instruments), onto 96-well sitting drop MRC 2 Lens Crystallisation UVXPO Microplates (SWISSCI), with a typical reservoir volume of 50 µl and a combined drop volume ranging between 0.4-1 µl (with varying ratios of protein to precipitant).

Post dispensing, plates were incubated and imaged in a ROCK IMAGER® (Formulatrix) either at 20 or 4 °C. Images were collected using visible, UV and cross polarised light and plates were monitored for crystal growth for up to one year. Any hits identified were either directly looped and frozen in liquid nitrogen (see **SECTION 2.9.1**) or an optimisation screen was setup around the hit conditions, altering factors such as pH, precipitant, protein concentration, drop ratio etc.

2.9.1 Crystal Looping and Data Collection

Promising crystals that we wanted to collect X-ray scattering data for were looped using a LithoLoop™ (Molecular Dimensions) with cognate dimensions to the crystal. The looped crystal was briefly incubated (< 2 min) in a cryoprotectant solution containing 30 % (v/v) glycerol, if there was not already a cryoprotectant at sufficient concentration in the drop solution. The cryoprotectant was a close condition match (identical buffer, pH, salts and precipitants) to the original drop, however containing the glycerol. After incubating the crystal for approximately 1 min in cryoprotectant, the crystal was transferred directly into liquid nitrogen and kept there until data was collected.

Data were collected on a synchrotron X-ray beamline (I04-1, Diamond Light Source) with 360° of rotation with 6 data frames collected per degree. Data reduction and analysis was performed using the iMOSFLM (151), CCP4 (152) and Phenix (153) program suites where the specific procedures are described in the relevant sections.

2.10 Small-angle X-ray Scattering

Protein samples were prepared and concentrated to 5-10 $mg \cdot ml^{-1}$ in a minimal buffer, typically 50 mM Tris-HCl pH 7.5, 200 mM NaCl. 45 µl of sample was loaded onto either a Superdex 200 Increase 3.2 (2.4 ml) or Shodex kW-403 (4.6 ml) column connected to an HPLC where the frames were captured on the eluate at a rate of 3 $f \cdot s^{-1}$. Data were collected on a high energy X-ray beamline at a synchrotron (B21, Diamond Light Source). The collected data was subsequently processed using ScÅtter (version 3.0). Frames corresponding to peaks on the $A_{280\,nm}$ trace were merged and averaged about regions where the estimated R_g was consistent. Buffer subtraction was performed using averaged frames directly prior to the first peak. Further SAXS analysis, including Guinier analysis, $P(r)$ distribution fitting and dummy atom modelling are described in detail in the appropriate section.

2.11 Molecular Dynamics

Spatiotemporal molecular dynamics (MD) was performed using GROMACS (Groningen Machine for Chemical Simulations; version 5.0) (154, 155). Simulations were conducted using either the GROMOS 54a7 (156) or the OPLS-AA/L (157) forcefields. The system was solvated using the SPC/E water model (extended simple point charge model) (158) prior to adding 0.1 M Na^+ and Cl^- ions (by replacing water molecules) and neutralised to a net charge sum of 0 with additional Na^+ or Cl^- ions as necessary.

The system underwent energy minimisation prior to temperature and pressure equilibration. Full details of the default molecular dynamics parameters (MDP) are given in the digital information appendix (**APPENDIX I**). The system was temperature and pressure equilibrated for 100 ps each before running 1-100 ns production MD. For longer time courses, MD was performed on the BlueCrystal high performance computing (HPC) machine utilising MPI to run across 100s of computer cores or in GPU-accelerated mode.

The resulting trajectory was re-centred on the protein while removing the periodic boundary conditions (PBC) that were set on initialisation. The root-mean-square deviation (RMSD) was then calculated over the MD time course on the protein backbone. Further details will be provided in the necessary sections.

2.12 Whole Genome Sequencing

Isolates chosen to be sequenced had their genomic DNA prepared using a DNeasy® Blood & Tissue kit (Qiagen) according to the manufacturer's instructions. Genomic DNA integrity was verified on a 0.8 % (w/v) Agarose Tris-Borate-EDTA (TBE) gel – any lanes containing a smear, indicating DNA degradation, were re-purified. Purified genomic DNA was sent to MicrobesNG™ (Birmingham, UK) who carried out sequencing, assembly and initial annotations. Genomes were sequenced with a minimum 30X coverage using the Illumina MiSeq platform with 250 bp paired-end reads.

2.13 Phylogenetics

2.13.1 Genome and Proteome Mining

All up-to-date (as of March 2018) genomes and proteomes tagged as *Fusobacterium* were downloaded from the GenBank and RefSeq FTP repositories (159, 160) using a search and retrieval script. Both databases were used as some files are only found in one and not the other. GCA (GenBank) and GCF (RefSeq) assembly accession numbers were filtered to use the GenBank over the RefSeq accession where available. Human-readable names were then assigned using the species and strain information found within the assembly report file.

2.13.2 Average Nucleotide Identity

Average nucleotide identity (ANI) scores were calculated using a program called orthoANI (161) which utilises USEARCH (162) to implement fast genome searches and comparisons. ANI scores, denoted A_{abs}, range from 25 % (all noise; meaningless result) to 100 % (complete identity). For use in some post-analyses, A_{abs} was converted to relative distance (A_{rel}) using **EQUATION 2.1**.

Equation 2.1 | **Relative Average Nucleotide Identity.**

$$A_{rel\,x} = \frac{1 - A_{abs\,x}}{1 - \min\,(A_{abs})}$$

When A_{rel} equals 0, the two strains are identical and a score of 100 is the most distant comparison in the full set. Because the relationship is increasing with distance, phylogenetic trees can easily be created using this value.

2.13.3 Maximal Unique Matches and MUMi

The MUMmer package was used to generate unique matches between two genomes. Genome sequences containing multiple scaffolds or contigs were concatenated prior to running MUMmer, such that the output result could be processed by downstream scripts

provided with the package. The input for MUMmer consisted of each genome with a minimal match length (l) equal to 19 to produce the following terminal command:

"mummer -mum -b -c -l 19 genome1 genome2 > mummer_out".

The MUMmer output file was then parsed to a Perl script (163) with the two genome lengths, which outputted a value between 0 and 1 for genomic distance: 0 being identical and 1 being extremely distant – this is called a Maximal Unique Match index (MUMi; M_i).

A script was written in Python to compare all downloaded genomes to one another utilising multiple processes. To increase speed and efficiency of the included MUMi script, it was translated into an optimised Python script that could directly process the MUMmer output. The scripts used are posted on GitHub (164).

To produce a M_i that correlated with A_{abs} and A_{rel} in a linear fashion, a transformation was applied to M_i in post analysis, with the result denoted M_l in **EQUATION 2.2.**

Equation 2.2 | **MUMi Linear transformation.**

$$M_l = \ln\left(\frac{1}{1 - M_i}\right)$$

Identical strains remain at 0, however, more distant strains will separate further, with no upper bound. For the strains tested, M_l positively correlates with A_{rel} with an approximate linear relationship. Like A_{rel} this score can also be used to create phylogenetic trees.

2.13.4 Pan-Locus Sequence Analysis

Conserved protein sequences were isolated from the whole proteome set of *Fusobacterium* using successive BLAST (165) queries and filtering out hits with % similarity < 65 % and coverage < 60 % to produce refined proteomes with ordered analogous protein sequences. From these, random gene samples (n = 10) were taken from each, concatenated and aligned using MUSCLE (166), with the maximum number of iterations set to 2 to limit the time spent aligning. Pairwise distances were calculated from the alignments using the

Poisson correction model in MEGA7 (167). The random sampling, alignment and distance calculation steps were repeated 1000 times.

The distances between strains for all trees were collated and averaged for each strain pair, denoted N_s (average number of amino acid substitutions per site). $N_s = 0$ is an exact match and N_s approaching 1 is very dissimilar – and impossible in the analysis, as dissimilar proteins were filtered out in the pre-processing stage. The scripts for this analysis can be found at GitHub (168). No post-processing was applied to this score as it already has the desired properties.

2.13.5 Phylogenetic Trees

Pairwise distances from N_s, M_l and A_{rel} were arranged into the MEGA or Nexus file formats, from which, phylogenetic trees were created using MEGA7 (167) or SplitsTree4 (169) respectively. These trees were then converted to either unrooted phylogenetic trees or rooted phylograms using the BioNJ (Biological Neighbour Joining) method (170, 171) or the UPGMA (Unweighted Pair Group Method with Arithmetic Mean) method as stated when performed.

Protein sequences were aligned using Clustal Omega (172) before creating phylogenetic trees in MEGA7 using the BioNJ method with 1000 bootstrap replications to measure branch confidence.

Group clustering analysis was performed using a simple agglomerative algorithm that grouped strains based on genomic distances in accordance to the similarity threshold that was set prior. The corresponding outgroups were then scored against each other using the mean distance between each individual pair within the two groups. From these scores, condensed phylogenetic trees could be created using the same method as described previously.

2.13.6 Hive Plots

For direct visualisation between two or more lengths of DNA, hive plots were used to show the regions of homology graphically. To do this, a Python script was written to perform homology searches between regions of DNA using BLASTN (version 2.7.0). In essence, the first length of DNA was split into sequential, overlapping, shorter chunks with equal lengths. The lengths of the chunks and the region of overlap varied according to the desired resolution, for example, a whole genome of about 2 Mbp would be split into 1000 bp regions with 500 nucleotide overlaps. A BLAST database was created for the second genome using the BLAST+ suite (165), before carrying out searches for each length of DNA. Up to four of the highest scoring hits were retained and evaluated based on sequence identity. A cut-off for identity was set, by the user, and any hit below this threshold was excluded from the graphical representation as well as regions that had no hits. Hits were then matched between two linear representations of the lengths of DNA and a connecting line was coloured according to the percent identity. The script for this application was posted on GitHub (173).

2.14 Statistics

Unless otherwise stated, all experiments conducted consisted of at least three biological repeats, including three technical replicates where practical. The relevant statistical tests are stated in the text, with p values lower than 0.05 considered significant. The R software package (174) was utilised to perform tests such as ANOVA (Analysis of Variance) or PPMCC (Pearson's Product Moment Correlation Coefficient). Phylogenetic statistics, unless otherwise stated, were generated in the MEGA7 application (167).

Chapter 3: A comprehensive phylogenetic analysis of the *Fusobacterium* genus

3.1 Introduction

Recently, conflicting classification of some *Fusobacterium* species has been reported in the literature (22, 23, 175) and it is now raising questions regarding phylogenetic boundaries in general. The current accepted definition of species, from a genomics perspective, is ≥ 70 % DNA-DNA Hybridisation (DDH) and/or ≥ 98.6 % 16S identity, though other more stringent values have been suggested (176).

The importance of describing a species is vital in many fields ranging from medicine to biosecurity, though none of these rely solely upon genomic similarity for the importance of classifications, rather use their shared phenotypic traits. However, it is likely that organisms that have closer genetic similarity will possess a similar phenotype, though this is excluding mobile genetic elements. In the case of *Fusobacterium* spp., mobile genetic elements are less concerning when examining genetic composition as they are not naturally transformable due to their diverse array of restriction endonuclease systems, though some strains have been found to possess plasmids (177-179). Therefore, by only examining the genome, this would provide enough information to determine the most likely phenotype of any given strain.

Modern computational approaches can reverse engineer DDH values, such as Average Nucleotide Identity (ANI), Genome-Genome Distance Calculator (GGDC) and Maximal Unique Matches index (MUMi). It has been shown that these methods correlate well with empirically derived DDH values, with a 95 % ANI score representing approximately 70 % DDH (180). Computational and empirical DDH values both exhibit pitfalls that over- and underestimate species difference, for example, neither method considers mobile genetic elements that may be passed around between a subset of organisms, though this is less of a concern for *Fusobacterium* spp., for reasons previously explained. Other intra-genome

rearrangements would decrease DDH considerably; however, this limitation is largely overcome by most computational methods that do not bias neighbouring genes as strongly. However, incomplete genomes from whole-genome sequencing (WGS) studies will reduce the accuracy of computational methods and the genus of *Fusobacterium* has many incomplete (contig/super-contig/scaffold levels) genomes. Further complications exist inherently within all *Fusobacterium* strains where the GC % content is extremely low, ranging between 25 and 35 % typically across the genus, which has the effect of biasing certain sequences, hence leading to overrepresentation of short runs of nucleotides.

It has been proposed that the *Fusobacterium nucleatum* (*Fn*) species be dismantled and the current acknowledged subspecies of *animalis, nucleatum*, *polymorphum*, *vincentii* and W1481 be reclassified as separate species (22). These proposals were justified using reverse engineered DDH values produced by ANI and GGDC methods. Some of the results, however, conflicted with 16S rDNA comparisons where the predicted DDH value was < 70 % and 16S rDNA was > 98.6 %. One such result came when comparing *Fn* subsp. *polymorphum* CTI-6 strain to the *F. hwasookii* KCOM 1249^T type strain, which exhibited a 99.7 % 16S identity and 49.3 % predicted DDH. As later discussed, the scope of this study could have been extended further into the realms of the whole genus, with the omission of the *F. periodonticum* species and no representation of the other clinically important strain, *F. necrophorum*, which both would have impacted the conclusions drawn.

This study explores different computational methods for classifying the entirety of successfully sequenced genomes (at the time of writing) from the *Fusobacterium* genus and what implications this may have on the taxonomy for the genus. Furthermore, some new, previously uncharacterised clinical strains will be sequenced and brought into the analyses. Three principle approaches were used for computational analysis, one proteomic method and two established genomic analyses respectively: Pan-Locus Sequence Analysis (PLSA), MUMmer (Maximal Unique Matches) and Average Nucleotide Identity (ANI).

In addition to the analyses conducted in this study, the results can be compared with a much larger recent global study regarding the reclassification of all bacterial species (23).

3.2 Whole-Genome Sequencing of Clinical Strains

For work carried out in later chapters of this study, we examined various clinical *Fusobacterium* isolates for their CEACAM1-bining propensities, and the proteins related to this. By comparing our strains to the current database, we aimed to use these additional data to validate the genetic comparison approaches used in this chapter.

From our lab collection of clinical strains, obtained from the Anaerobe Reference Unit (Cardiff, UK; **TABLE 2.1**) and School of Oral and Dental Sciences (Bristol, UK; **TABLE 2.1**), 25 isolates were chosen for whole-genome sequencing. The strains to sequence were primarily chosen for whether they could bind CEACAM1, as well as a few that did not to for comparison. The CEACAM1 screening is described further in **CHAPTER 4**. Additionally, we aimed to keep the associated pathologies from which the strains were cultured as varied as possible. Four strains were isolated from blood culture, five from the oral cavity, two from cerebrospinal fluid and the remaining from various abscesses and pus.

The whole-genome sequencing method is described in detail in the Methods (**SECTION 2.12**). In short, the bacteria were grown under anaerobic conditions and had their genomic content extracted and purified. This was then sent off for sequencing, where library preparation, sequencing and initial analyses were conducted. Draft genome assemblies were returned together with gene annotations and data quality. All the strains had a genome coverage that exceeded 30X giving sufficient depth for further analyses.

From the results, 22 of the strains examined returned whole genomes at the contig assembly level. Each possessed genomic traits expected of *Fusobacterium*, such as approximately 2 million base pairs (Mbp) of DNA and a GC % content between 25 and 35 %. Three of the tested strains exceeded these expected parameters. The R33458 strain was found to be a mixed culture of *Fusobacterium* (likely *vincentii*) and *Bacillus* sp.

(unidentified species). The 2B17 strain had an abnormally large genome (3.7 Mbp), though none of the genomic content appeared abnormal with respect to *Fusobacterium*. This could have been due to poor assembly or issues with library preparation; however, the genome could still be used in some downstream analyses. The 2B3 strain also could not be resolved to a single species and finished with a large heterogeneous genome of 8.2 Mbp, and thus could not be used accurately within ANI and MUMi analyses as there is a bias introduced with any contaminating genomic content. However, for all 25 strains examined, PLSA could still be carried out, as any anomalous genes from contaminating organisms were discarded at the filtering stage (see PLSA method; **SECTION 2.13.4**).

Strains were assigned to a specific species/subspecies using MUMi against the reference type strains that have been identified. For 5 strains, excluding the aforementioned anomalous strains, a subspecies/species could not be matched. For the remaining strains, they each fell into either the *nucleatum*, *vincentii* or *animalis* subspecies, with no representation of *polymorphum*, W1481, *hwasookii* or *periodonticum*. **TABLE 3.1** lists the results of the WGS together with the initial classifications made.

The reference strains used were: *Fnn* ATCC 25586^T (*nucleatum*), *Fnv* ATCC 49256^T (*vincentii*), *Fna* 7_1 (*animalis*), *Fnp* ATCC 10953^T (*polymorphum*), *Fnw* (W1481), *Fperio* ATCC 33693 (*periodonticum*), *Fh* ChDC F128 (*hwasookii*), *Fmal* Marseille-P2749 (*massiliense*), and *Fr* ATCC 25533 593A (*russii*). A MUMi score ($M_i(19)$) less than 0.4 considered the same species (the reason for this value is explained in **SECTION 3.4**). As all strains examined had initially been typed as *F. nucleatum* originally, only strains within the major clade surrounding *nucleatum* were initially used for testing, with the most distant strain used being *F. russii*.

Table 3.1 | **Whole-Genome Sequencing Results.**

[1] Species prediction compared the MUMi result to a representative of each species. Species names listed conform to the new nomenclature explained later. [2] No matched type-strain, so assigned as novel species. [3] Inconclusive sequencing. [4] Low coverage. 2B- strains from SODS. R- strains from ARU.

Strain ID	N Contigs	N50 (bp)	Coverage (X)	Total Length (bp)	Species Prediction[1]
2B16	27	218887	65.5	2248027	nov.[2]
2B17[3]	967	52747	34.1	3808046	*vincentii*
2B2	106	44814	36.2	2279829	nov.
2B3[3]	891	33561	59.8	8247164	multiple
2B4	18	282219	235.8	2261255	nov.
R15792	135	49336	136.0	2640740	*animalis*
R16531	109	44665	43.2	2122526	nov.
R18528	80	71186	117.5	2457181	*nucleatum*
R18932	91	52467	48.3	2291618	*animalis*
R24394	223	18552	33.2	2268759	*nucleatum*
R26872	18	369250	118.6	2251292	*vincentii*
R28211	76	71725	56.4	2343767	*vincentii*
R28385	107	36920	32.6	2136526	*nucleatum*
R28400	68	60923	56.0	2221750	*nucleatum*
R28427	45	69838	111.3	2152299	nov.
R29976	64	66844	30.7	2191054	*vincentii*
R30464	21	503369	270.2	2098906	*vincentii*
R30604	33	173259	106.8	2197183	*vincentii*
R30927	166	41066	84.3	2784735	*animalis*
R31249	23	259872	111.3	2095525	*vincentii*
R32310	27	216428	52.5	2237187	*vincentii*
R32935	56	99708	48.1	2178570	*nucleatum*
R33458[3]	903	30252	27.3[4]	8140144	multiple
R33533	48	96764	41.4	2140227	*vincentii*
R5001	229	19824	35.8	2574309	*animalis*

3.3 Method Comparison

3.3.1 Pre-screening for Non-*Fusobacterium* Strains

Prior to continuing with other analyses, certain strains were found to most likely not be *Fusobacterium*, though their classification suggested otherwise. These include *F.* sp. CAG:815, *F.* sp. CAG:439 and *F. naviforme* ATCC 25832. These strains were likely classified according to 16S rRNA identity, which is not a good metric when comparing *Fusobacterium* species as it can be often misidentified for *Clostridium* 16S rRNA and, as previously mentioned, distinct species within *Fusobacterium* have 16S % identities higher than what is needed for species definition. These discrepancies were found in preliminary testing with MUMi against representative strains from each species. These three strains all had a score distinctly above any other pairwise comparison, with *F. naviforme* ATCC 25832 being the most dissimilar to everything else. It is overwhelming likely they are not *Fusobacterium* strains – the fact they lack the *mre* genes (rod-shape determining proteins; highly conserved among all *Fusobacterium*) immediately gave cause for concern. The *mre* genes by themselves are a good indication for classifying a bacterium within the *Fusobacterium* genus, however comparative resolution is lost when examining species/subspecies relationships, which is a similar drawback when using 16S rRNA-based classification.

We also included the newly sequenced strains in all analyses conducted, partly to confirm the initial classifications made, but also whether comparing them to all strains in the database, we could identify the unclassified strains or even classify them as novel species. Three clinical strains, R33458, 2B17 and 2B3, were excluded from analysis with MUMi and ANI due to likely contaminating genomic content; however, they were still included in PLSA as this should not be greatly affected.

3.3.2 Average Nucleotide Identity

The first approach used for whole-genome analysis was Average Nucleotide Identity (ANI), which is the current gold standard for computational genomic distance calculation (180).

This yields an overall percent identity between two organisms. A pair scoring above 95 % are generally considered matching species, based on a reversed engineered DDH value of approximately 70 % (180). For each strain in *Fusobacterium*, a pairwise ANI score (A_{abs}) was given to each using the orthoANI software package (161). This leverages the USEARCH program (162) for performing genomic searches and uses the results to infer overall identity. This provides a severe limitation in computation speed and can take days or even weeks of computation time for only a few thousand pairwise comparisons. This program had to be compiled and run on the inhouse supercomputer BlueCrystal (University of Bristol) in order to achieve results more quickly. Aside from the computational difficulties, ANI provided the most human-readable result in clear genomic percentages, where other methods were somewhat more abstract. In addition, the species threshold for this technique has been previously established (180).

3.3.3 MUMmer and MUMi

The second genomic analytical method used the MUMmer (Maximal Unique Matches) application with the related MUMi (MUM index) package (163, 181). This method uses MUMmer to identify unique matches with a minimum length between genomes and then lists the regions, which is then fed to a script (MUMi) that produces a score. The pairwise scores (M_i) produced are distilled to a value between 0 and 1, with 0 being identical and 1 being very distant (see Methods for details; **SECTION 2.13.3**). Throughout the remainder of this study, the minimum match length for MUMmer, l, was set to 19. This was determined to be a suitable value for whole-genome comparisons between *Fusobacterium* strains and was used in a previous study comparing *Fusobacterium* species/subspecies (175). Higher values would increase resolution for similar strain comparisons while decreasing the resolution for more distant pairs.

To improve the execution speed and compatibility of this program, the MUMi-generating script as described in (163) was converted into a Python script that could directly process the MUMmer output in a more efficient manner. All the scripts were then packaged into one

main script, which again allowed for the leveraging of multiple processors to improve speed and scalability. The scripts used are posted on GitHub (164).

One limitation of MUMi is that comparable resolution is lost when examining two more distant genomes. This is primarily an issue for data visualisation as it is harder to interpret the scale with increased resolution for more similar organisms. To overcome this issue, the MUMi value was transformed (**EQUATION 2.2**) such that a linear correlation to ANI was produced (the transformed value is described as M_l; see Methods), thus increasing resolution at the distant ends. **FIGURE 3.1** shows all genome comparisons using M_i and M_l values and how they differ graphically and their correlations with respect to ANI and PLSA.

Formulae representing MUMi will represented in the following form herein: $M_x(l)$ where l equals the minimum match length and x equals either i (raw) or l (transformed) depending on the scenario.

3.3.4 Development of Pan-Locus Sequence Analysis

In the past, MLSA (Multi-locus Sequence Analysis) based studies used a very limited subset of genes (typically fewer than 10) known to be found within all target species to classify different strains based on more subtle changes in sequence. In this method, all similar genes were identified within all species of *Fusobacterium* using a proteomic-based search with a cut-off value of > 65 % similarity and > 60 % coverage. This was achieved using an iterative local BLASTP (BLAST+ version 2.7.1) search with filtering and refinement at each stage, the Python script used (168) was optimised for multicore execution running multiple BLAST searches split across all cores before refining and running new restrained searches. This method vastly increased execution time versus optimising the individual searches themselves across many cores, as this did not fully utilise the performance available.

In total, 215 conserved protein sequences were found, out of approximately 2000 total predicted protein-coding regions per organism. The number would almost certainly be higher if more of the deposited sequences in GenBank (182) and other genome archives

were complete, therefore reducing the number of partial genes that would be discarded at the filtering stage due to a lack of coverage. This is an inherent problem with Illumina-based sequencing platforms where read lengths are shorter and so the resulting assembly has more short gaps where repeat sequences for example could exist. As confirmed with MUMi in the preliminary stage, two labelled *Fusobacterium* strains, *F.* sp. CAG:439 and *F.* sp. CAG:815, were so different from everything else that the conserved gene pool produced was very limited at fewer than 50 and so were omitted from the analysis.

From the refined list of conserved proteins, a 10-protein random sample was taken, concatenated and aligned using MUSCLE (version 3.8.31). The sample size was set at 10 to ensure enough internal diversity, but also not too many such that execution time was greatly increased. Pairwise distances were calculated from the alignment using the Poisson correction model (183) in MEGA7, which produced a result pertaining to the average number of substitutions per site. This process was repeated 1000 times and the distances between each strain pair were averaged. As before each individual alignment was assigned to a single thread where multiple alignments and scoring was performed across as many cores as available, as opposed to running MUSCLE in multithreaded mode.

A script was written in Python to combine all the stages of the process of the analysis and completely automate execution for many strains (168).

3.3.5 Score Comparison

Overall, all three methods yielded data that strongly correlated with one another. When a Pearson's Product Moment Correlation Coefficient (PPMCC) is applied the following statistics were obtained: $\rho \geq 0.98$ or $\rho \leq -0.98$ and $p < 0.01$ for all comparisons of ANI, LMUMi and PLSA (**FIGURE 3.1**). **FIGURE 3.2** shows the comparison MUMi and PLSA against ANI at the highly similar end of the scale demonstrating where the cut-off for species definition is for each method. From looking at these data, it is very clear where the cut-off for ANI lies as there is a distinct separation at the 94 % mark; however, the same cannot be attributed to PLSA or MUMi, though the species cut-off for MUMi groups the same strains

as ANI does. This is mostly the case for PLSA, though, there are a number of conflicting results, but these are restricted to *F. periodonticum* comparisons which will be discussed later.

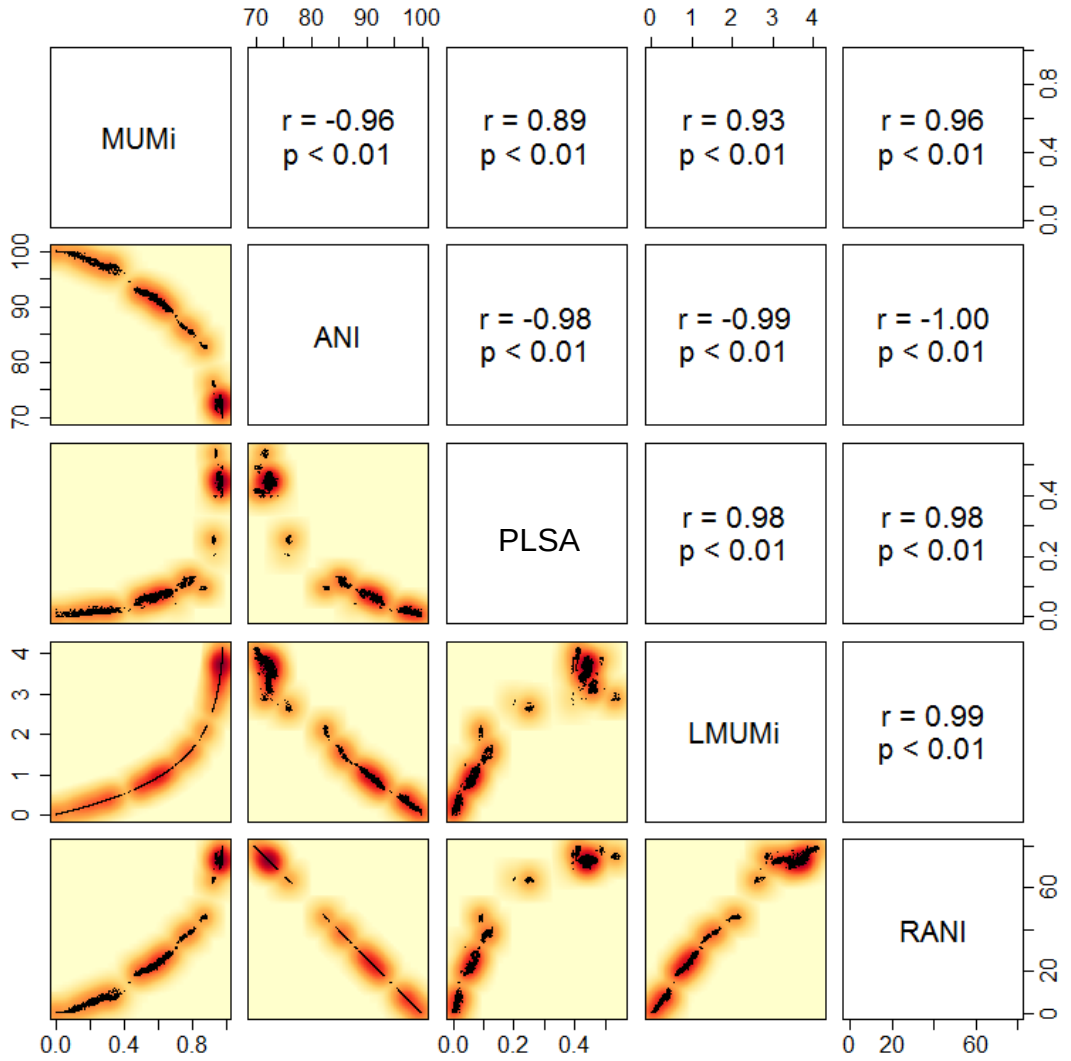


Figure 3.1 | **Distance method overall correlations.**

The above set of graphs show the correlations between the three methods used with r (ρ) and p values according to Pearson's Product-Moment Correlation test. Included are two transformations applied to ANI and MUMi (**EQUATION 2.1** and **EQUATION 2.2** respectively) that were used in some analyses. All methods correlate strongly, though variance increases at the low-resolution comparisons. MUMi – Maximal Unique Matches index; ANI – Average Nucleotide Identity; PLSA – Pan-Locus Sequence Analysis; LMUMi; Linear-transform of MUMi; RANI – Relative-inverted ANI.

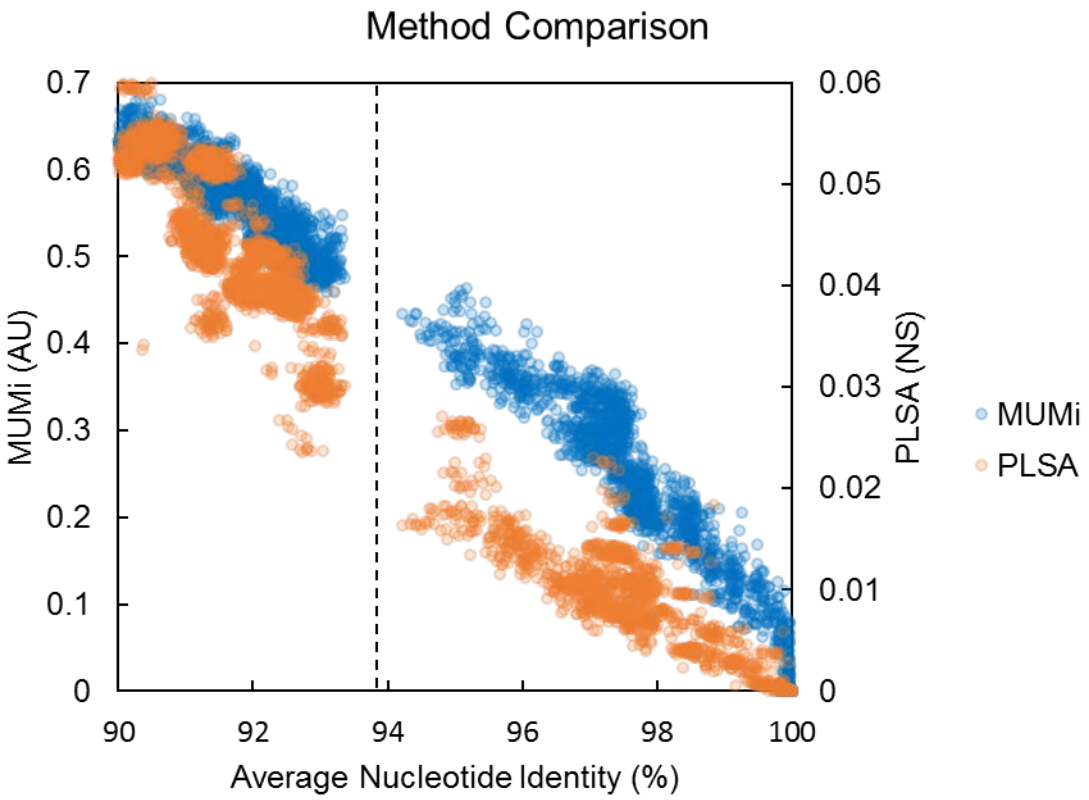


Figure 3.2 | **Comparison between the three methods used at the intra- and interspecies interface.**

This species interface can be clearly identified from examining the ANI score, where there is a distinct gap in the data. This gap is not so evident when only looking at the MUMi or PLSA results; however, MUMi does have the same gap as ANI, though it is much slighter. PLSA on the other hand does not share this gap and there are a group of conflicting results. ANI v MUMi – blue; ANI v PLSA – orange.

Each method has their own advantages and disadvantages for speciation and it is not entirely certain if only performing one analysis is sufficient. Nonetheless, each of these methods has a distinct improvement over 16S rRNA sequence identity or single gene comparisons, by taking direct observations spanning entire genomes.

3.4 Defining the Taxonomic Boundary

For the majority of strain comparisons, a clear definition is seen between intraspecies and interspecies in agreement with Kook *et al* (22). The highest score between two distinct

species came from *F. hwasookii* ChDC F206 and *Fn* subsp. *polymorphum* KCOM 1275 at $A_{abs} = 93.37\,\%$. This result compares with findings by Kook *et al* (22) and shows *F. hwasookii* is a separate species. From this point, strain comparisons with $A_{abs} < 94\,\%$ can be considered different species.

For MUMi, the cut-off value can be set at $M_i(19) = 0.456$. The intraspecies scores close to this value are the same that go below the 95 % ANI score. MUMi has previously been used to determine that the W1481 strain belongs to the *F. nucleatum* species (21). This was determined under the assumption that all subspecies under *Fn* belong to the *nucleatum* species, however, it is evident that this is not the case. To contain W1481 within *Fn* (all strains within the old species) a threshold of $M_i(19) = 0.693$ ($A_{abs} \approx 89\,\%$) is required – this would mean that *F. hwasookii* species would also belong to *Fn*, moreover it causes *F. varium* and *F. ulcerans* to be within a single species. As this is not the case, where these species are distinct, in agreement with Kook *et al* (22), the previous subspecies of *Fnp*, *Fna*, *Fnn*, *Fnv,* and *Fnw* should be classified as their own individual species.

For PLSA, the corresponding species threshold lies at 0.0232 average number of substitutions per site across the 215 conserved proteins. This boundary is adhered to by the large majority of intraspecies comparisons, with the only exception being for *F. periodonticum* ATCC 33656 and the rest of *F. periodonticum* which will be discussed later. PLSA becomes less well correlated with MUMi and ANI at the more distant ends of the comparisons, as seen in **FIGURE 3.1**, where the clusters are much larger. There are also visible clusters throughout the complete range, where ANI or MUMi rate the differences either more or less stringently than PLSA. This could be attributed to silent mutations and differences in the non-proteomic content, as the genomic content represented by the 215 conserved genes is approximately a tenth of the total content. Nevertheless, PLSA may be advantageous, as MUMi and ANI do not consider silent mutations or mobile genetic elements that can be lost or gained, affecting overall identity. When PLSA was restricted to

a single species, more proteomic content could be retained, therefore increasing the accuracy of the analysis.

There are a number of genome pairs that fall close to the species boundary and below the previous threshold of 95 % average nucleotide identity. These pairs all originate from the *F. periodonticum* and *F. polymorphum* species, for example *F. periodonticum* 1_1_41 and *F. periodonticum* D10 have an ANI score of 94.43 %, though comparisons with other *periodonticum* strains are higher than the cut-off point, thus representing a diverse species with a broad spectrum of strains. See **TABLE 3.2** for the fringe cases and the point at which interspecies comparisons are made.

Table 3.2 | ***Fusobacterium*** **comparison fringe cases.**

This shows the interchange between species and non-species comparisons, from *periodonticum-periodonticum* comparisons and other closely related species that fall outside the classification window, for example *polymorphum-hwasookii*. The table is sorted by ANI score and the three metrics used are given, as well as the accessions (where applicable) and the predicted species. Key: *perio – F. periodonticum*; *poly – F. polymorphum*; *hwas – F. hwasookii*; *anim – F. animalis*; *vinc – F. vincentii*; nov. – unidentified species.

Accession 1	Accession 2	ANI	MUMi	PLSA	Species	
GCA_000163935.1	GCA_000297655.1	94.43	0.434	0.0165	*perio*	*perio*
GCA_000163935.1	GCA_002763695.1	94.40	0.434	0.0163	*perio*	*perio*
GCA_000163935.1	GCA_002763925.1	94.37	0.425	0.0169	*perio*	*perio*
GCA_000163935.1	GCA_002761935.1	94.32	0.427	0.0164	*perio*	*perio*
GCA_000163935.1	GCA_002763915.1	94.22	0.434	0.0163	*perio*	*perio*
GCA_002211625.1	GCA_000455905.1	93.37	0.475	0.0301	*poly*	*hwas*
GCA_000479225.1	R16531	93.33	0.521	0.0353	*anim*	nov.
GCA_001455085.1	GCA_002211625.1	93.33	0.479	0.0301	*hwas*	*poly*
R16531	R30927	93.32	0.548	0.0350	nov.	*anim*
GCA_002749995.1	R16531	93.31	0.496	0.0366	*vinc*	nov.
R15792	R16531	93.30	0.530	0.0351	*anim*	nov.
GCA_002202115.1	GCA_000455925.1	93.30	0.481	0.0302	*poly*	*hwas*

The genus also appears to be very polarised with distinct major clades spread far apart with clusters of closely matched pairs extremely distant from other clusters. The *F. necrophorum* clade is furthest from the main cluster of the *F. nucleatum* with few organisms bridging the gap. Other non-typed strains can be seen to fall into groups, such as *Fusobacterium* CM21 belonging to *F. vincentii*. Numerous other loosely typed strains could be identified as specific species/subspecies (see **TABLE S 2** for details of classifications).

As there is no well-defined genus boundary at the time of writing, all scores below the species cut-off threshold bear little significance for the meantime, though there are definite clusters of scores around certain values. Approximately 30 % of comparisons have ANI scores > 80 and < 94 %, which allows the data to be split into three groups with intraspecies, similar species and very different species. The bounding thresholds for these groups can be set as follows: A_{abs} > 94, 80 < A_{abs} < 94 and A_{abs} < 75 % respectively.

3.5 Reclassification of Multiple Species

By defining a more rigorous set of parameters for classifying species, multiple reclassifications were made across the whole *Fusobacterium* genus. The full list of classifications if provided in **APPENDIX E**.

3.5.1 Identification and Classification of *F. oralis* sp. nov.

Using the defined taxonomic boundaries and novel data gathered through whole-genome sequencing, a new species emerges. This species lies on the *periodonticum*-containing branch at the split between the previous *Fn* and *Fperio* species (**FIGURE 3.3**). At least four of the clinical strains (2B2, 2B4, 2B16 and R28427) belong to this group with a fifth (2B3) likely but undetermined due to inconclusive sequencing results. One unclassified strain (*F.* sp. oral taxon 370 strain F0437) also belongs to this group. We suggest the name *F. oralis* sp. nov. for this species as every strain isolated originated from the oral cavity. This species has a genome length between 2.1 and 2.3 Mbp with a GC % content of 27.4 – 27.6 %.

This species currently is only associated with periodontal disease and related diseases – none of these species have been identified in other pathologies, such as colorectal cancer, preterm births or Lemierre's syndrome for example. That puts these bacteria very close to *F. periodonticum* in that respect, limited to oral infections. However, as there are currently very few strains that fall into this clade, it does not necessarily mean that this species is not associated with any other diseases.

3.5.2 Identification and Classification of *F. ovarium* sp. nov.

One of the clinical strains, R16531, did not fall into any species and lied almost equally between *Fn*, *Fv* and *Fa* (**FIGURE 3.3**). This strain was isolated from an ovarian mass; therefore, the name *ovarium* is suggested, from the Latin *ovarium inferum*. Very little is known about this strain as it the first of its type to be identified and may even represent a bridging strain that has not succumbed to the usual evolutionary trait of convergent evolution, whereby species tend to belong to a larger containing group.

As this is a singleton, little can be said about this species, other than what can be inferred directly from the data there is. It has a GC % content of 26.9 % and its genome is approximately 2.1 Mbp. Moreover, because it is a completely isolated case, disease relationship cannot be confirmed, though it has been shown that it existed within an ovarian mass.

3.5.3 Reclassification of various minor species

In accordance with Kook *et al* (22), the previously named *Fn* W1481 strain is actually further removed from the majority of other previous *Fn* species than, for example, *F. hwasookii* is from *polymorphum*. Its MUMi, ANI and PLSA scores are all far beyond the species threshold to its nearest neighbour – *F. animalis* (assembly accession: GCA_000220825.1; $A_{abs} = 90.21$; $M_i(19) = 0.639$; $N_s = 0.0553$). Once more, as this is an isolated case, no further conclusions can be made about this strain, but like with *F. ovarium* sp. nov., it may represent another strain that has diverged from the most commonly isolated lineages.

There are several other species, away from the old *Fn* and *Fperio* clade that have been incorrectly classified or have not been attributed to any species at all. One strain, assembly accession GCA_900015295.1, falls on a distant branch, shared with *F. mortiferum*. It is however, far removed even from this strain (also a singleton) and belongs to its own species. As it was originally classified as a *clostridium* species, evident on the name given to the specific strain, it will now be referred to as *F. closii* sp. nov.

The strain *F. equinum* (*Fe*) was previously identified and classified by Dorsch *et al* (184) and has remained uncharacterised ever since. This is contrary to the observations made in this study, where it is closely related to *F. gonidiaformans* (*Fg*), so much so, that is more closely related to *Fg* 3_1_5R than *Fg* ATCC 25563 is to the other *Fg* strain, with ANI and MUMi equal to 98.8 % and 0.190 for *Fe* v *Fg* 3_1_5R compared to 98.6 % and 0.215 for *Fg* 3_1_5R v *Fg* ATCC 25563. As there are two *F. gonidiaformans* strains and one *equinum*, it is therefore simpler to reclassify the one strain than the two, hence, *F. equinum* CMW8396 becomes *F. gonidiaformans* CMW8396.

Recently, a new *F. varium* strain was identified (185), however, the results of the work conducted in this study do not align with this classification. One key point is that it has a genome length of > 3 Mbp, which is far removed from what is expected across the entire genus, which suggests there is either a contaminating strain or it is not *Fusobacterium*. It is more likely the prior as its nearest neighbour, using MUMi and ANI, remains *F. varium*, though it is not that much further removed from *F. ulcerans* (*Fu*). PLSA, which should remove, at least non-*Fusobacterium*, contaminating genes still classifies it outside the boundary of *F. varium* and is likely its own species. The closest neighbour comparisons are as follows: $A_{abs} = 88.8\,\%$; $M_i = 0.695$; $N_s = 0.0316$ (compared against *Fu* ATCC 49185). Further work will be needed to confirm the true identity of this strain.

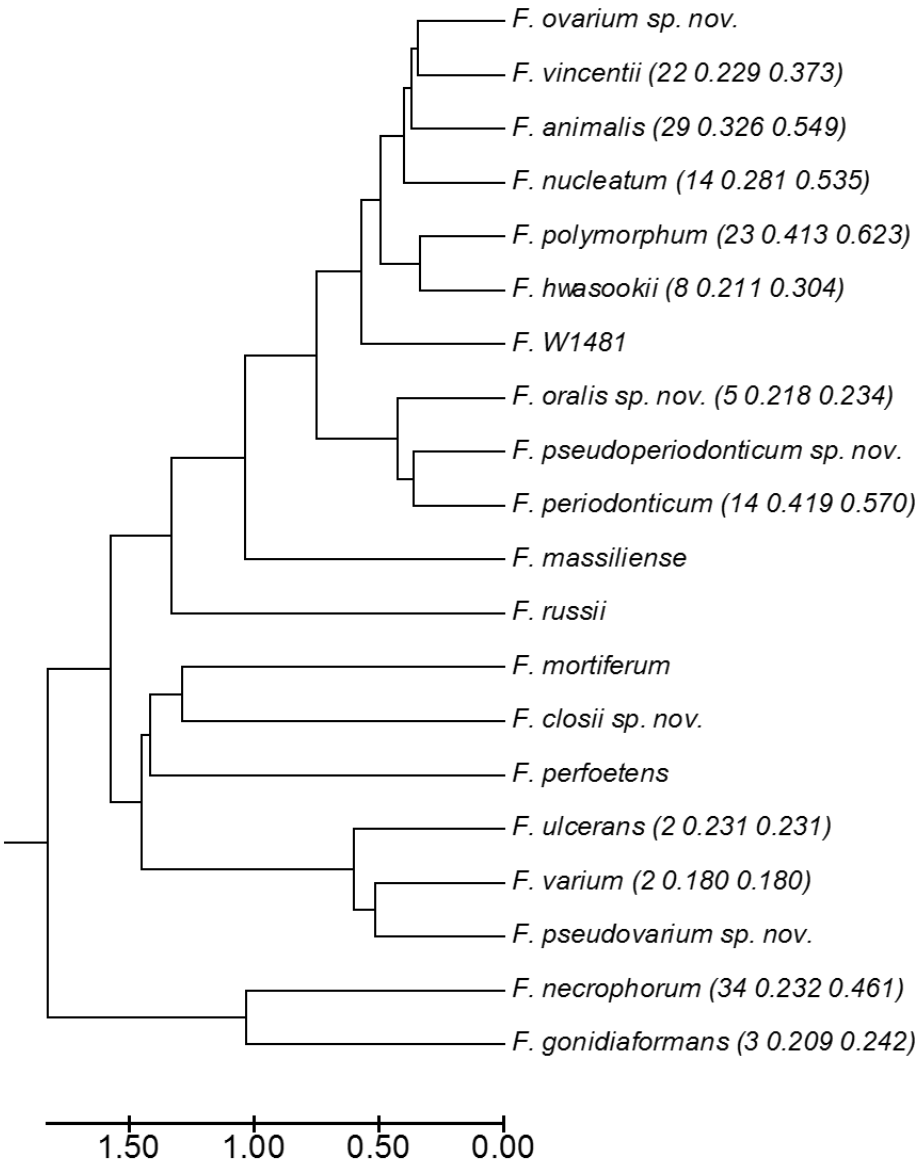


Figure 3.3 | **Phylogenetic tree for clustered species.**

All strains were clustered using an agglomerative algorithm to the species level based on the linear MUMi, M_l, metric. The bounding threshold for the clusters was set at $M_l(19) < 0.6$ which is equivalent to $M_i(19) = 0.451$. From this the evolutionary history was inferred using the UPGMA method in the MEGA7 application (167). The number of strains within the cluster, mean and maximum score are shown in parentheses after the species name for non-singleton species. The mean score gives a good representation of the spread for each species where a higher mean indicates a more diverse species.

3.5.4 The *F. periodonticum* paradox

For *F. periodonticum*, the type strain for this species is *F. periodonticum* ATCC 33656. This, interestingly, has one major problem in that under ANI and MUMi, this strain falls outside the definition of a species with ANI and MUMi scores equal to 93.0 % and 0.502 respectively to its closest neighbour (*Fperio* KCOM 1282). Examining the PLSA data, it falls within the species boundary definition, though existing on the fringes with a range of 0.0235 to 0.0267 for the whole of *F. periodonticum*. This compares to the largest intraspecies comparison of 0.0270 seen between *Fn* R18528 and *Fn* R28400.

As PLSA uses an algorithm that discards non-conserved or non-complete genes from the whole genus, any subtle differences in locally conserved genes within a single species by itself, could be missed. By performing PLSA for an individual species, fewer genes are discarded in the filtering stage, leading to a wider, more diverse gene set, which should increase the comparable resolution of sequence diversity. Running PLSA on the species of *periodonticum* (including ATCC 33656) and *oralis*, 1357 protein sequences are retained in the analysis. When comparing the ATCC 33656 strain to all other *periodonticum* strains, the minimum and maximum diversities become 0.0546 and 0.0580 respectively. The minimum value is 0.013 higher than the maximum intra-*periodonticum* comparison and is closer to a comparison to *oralis* at 0.0632. These scores cannot be compared directly to the original PLSA analysis as there are many more genes that are less highly conserved than the original 215 used.

Nevertheless, the fact that the genomic pairwise comparisons put this strain further from the main cluster of *Fperio* than *F. polymorphum* is to *F. hwasookii* for instance, raises the question whether this species even belongs to *Fperio*. As all other *Fperio* strains fall within the accepted species definition of one another, then it is more sensible to remove this single strain from the species, rather than reclassifying the remaining strains. The name *F. pseudoperiodonticum* sp. nov. (meaning false *periodonticum*) will be applied to this

particular strain for the remainder of this study. Of the 188 total strains compared (including the two non-*Fusobacterium* strains), this was the only anomaly of this kind.

3.6 Genomic Visualisation of WGS Strains

In addition to giving specific metrics to pairwise comparisons, to give a more tangible description of genomic similarity, genomic hive plots were created. Genomic hive plots are maps that connect two axes, representing two genomes, with a line representing a region of homology where a corresponding percent identity colour applied to each connection. Using these plots, it is easier to visualise whole genomes compared to each other, where features such as inversions, gaps and regions of low homology can be quickly identified.

To create these maps, a Python script was used that leveraged the BLAST+ program suite and the SVG writing library. See Methods for complete details (**SECTION 2.13.6**). In short, genomes had their contigs re-indexed against a reference genome, usually a complete genome from their parent species. The reordered contigs were split into short 2000 bp sub-sequences and these were mapped against the target strain using BLAST+. Hit lengths shorter than 1500 bp with less than 90 % identity were filtered out before connecting the axes. The stringent filtering increased execution speed and clarity of the resulting map. Other more specialised specifications can be set.

Genomic hive plots were created for 23 of the WGS strains using the following reference genomes for contig indexing: *Fn* ATCC 25586 (for *nucleatum*), *Fn* 3_1_36A2 (for *vincentii*), *Fn* 7_1 (for *animalis* and *ovarium*) and 2B16 indexed against *Fn* ATCC 25586 (for *oralis*). An example of unindexed contigs versus post-indexed contigs is displayed below in **FIGURE 3.4**.

These maps complement the raw genome-genome distance calculations providing a tangible representation of what each method is calculating. **FIGURE 3.5** gives another example of the comparisons that can be made. In this instance, it is displaying the 23 newly sequenced strains compared in a pairwise fashion to *Fn* ATCC 25586. The resulting maps

show the differences when comparing more distant strains to one another where the *oralis* comparisons produce a much thinner graph, whereas the *nucleatum* strains produce highly similar, densely packed maps with far fewer large-scale rearrangements.

FIGURE 3.6 shows four hive plots clustered for the sequenced strains ordered by species. Here, the species identities relative to themselves can be clearly observed, though *F. ovarium* sp. nov. R16531 had to be grouped with the *F. animalis* species as it is the only species of its type.

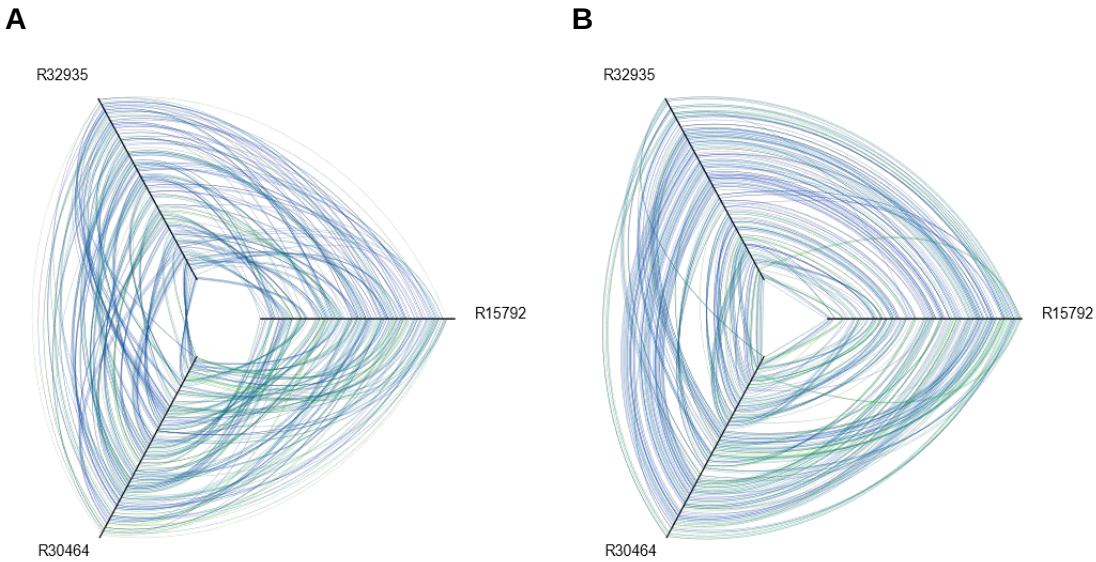


Figure 3.4 | **Unindexed contig hive plot versus indexed.**

A) Unindexed genomes of WGS strains R32935, R15792 and R30604. There are many regions mapping seemingly arbitrarily as a consequence of the contigs ordered by size within the genome. **B)** Contigs were indexed against *Fn* ATCC 25586 before comparing against each other, hence producing a clearer map.

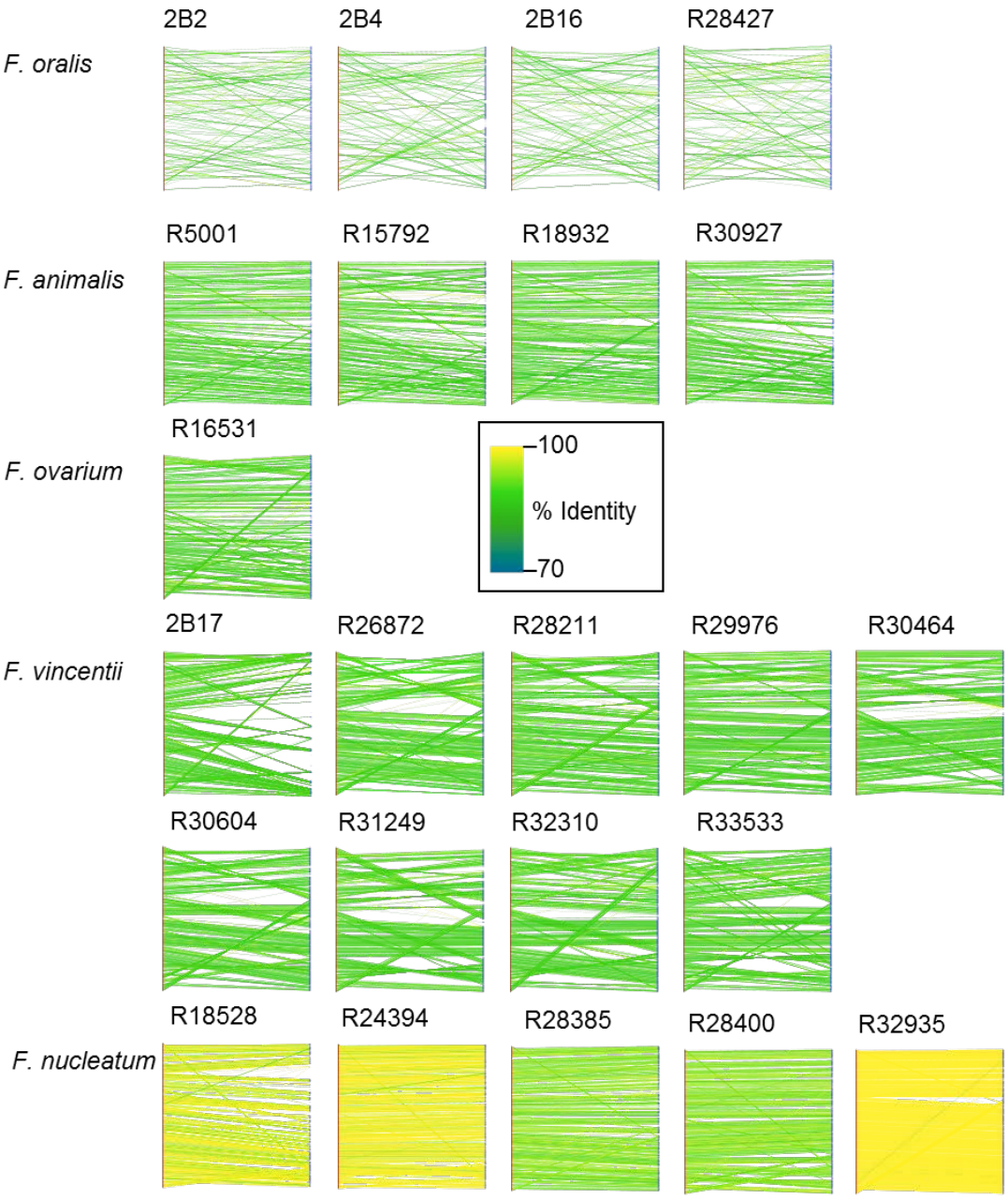


Figure 3.5 | **Pairwise genome maps for newly sequenced strains against *F. nucleatum* ATCC 25586.**

Contigs from all WGS strains (not including 2B3 and R33458) were ordered against their position in the reference strain *Fn* ATCC 25586. Positions of close identity were then connected with a line from the reference strain (left) to the target strain (right) and low scoring (hit length and % identity) regions were filtered for clarity. The maps were then grouped based on their predicted species.

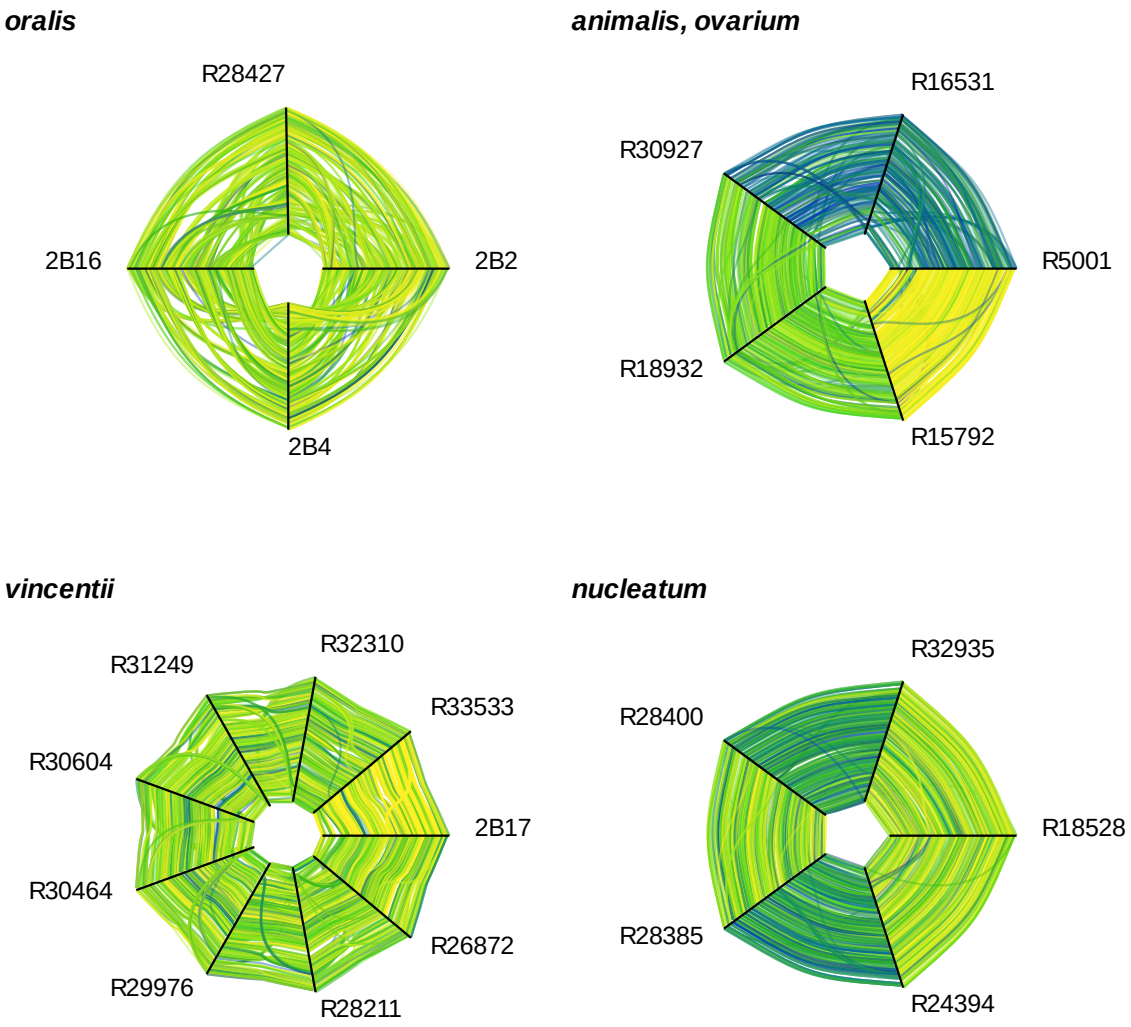


Figure 3.6 | **Hive plots for all sequenced strains grouped by species.**

The four genomic hive plots shown represent 23 genomes of the newly sequenced strains. They are correspondingly arranged by species group, with the singleton *F. ovarium* sp. nov. R16531 grouped with *F. animalis*. Strains for *F. oralis* sp. nov. do not have a completed reference genome so contigs were ordered against 2B16, itself indexed against *Fn* ATCC 25586.

3.7 Extension of Analyses to the Family

In addition to examining the *Fusobacterium* genus, a brief study into the *Fusobacteriaceae* family was conducted, which looked at whole-genome comparisons using MUMi. In addition to the strains used previously, 7 other taxa from *Fusobacteriaceae* were used: *Ilyobacter polytropus* (186), *Psychrilyobacter atlanticus* (187), *Fusobacteriaceae* bacterium UBA2433 (188), *Cetobacterium somerae* (189), *Cetobacterium ceti* (190), *Cetobacterium* sp. ZOR0034 and *Cetobacterium* sp. ZWU0022.

Pairwise MUMi scores were calculated for all strain pairs with the additional non-*Fusobacterium* species. The scores were transformed to the LMUMi metric (*see Methods*) and clades were collapsed to the species level using an agglomerative algorithm with a threshold set at $M_l(19) < 0.6$. The evolutionary hierarchy was inferred from the mean clade differences using the UPGMA method. The resulting phylogenetic tree is shown in **FIGURE 3.7**. From this, it is evident that *F. necrophorum* is as far removed from the *F. nucleatum* branch as other non-*Fusobacterium* strains such as *C. ceti.* In addition to the pairwise genomic distance scores, the taxa were arranged in a way such that the GC % content was increasing going down the tree, which for the most part is the case, except for *F. perfoetens*, which has a GC % content of 26.0 %, which is not comparable to its nearest neighbours of *F. mortiferum* (29.0 %) and *F. closii* sp. nov. (28.9 %).

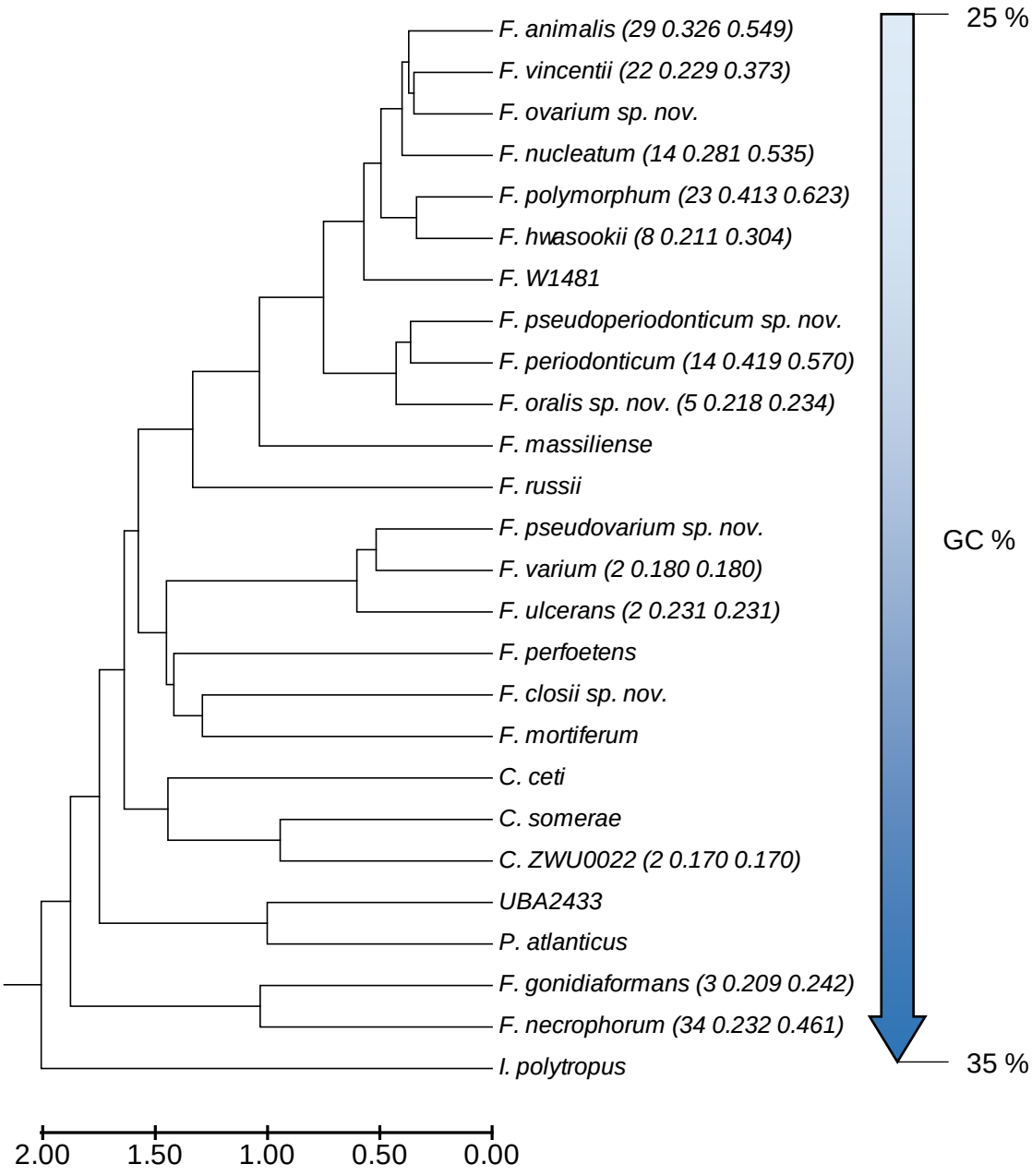


Figure 3.7 | **Phylogenetic tree of the *Fusobacteriaceae* family.**

The above phylogenetic tree was constructed using UPGMA evolutionary inference of grouped taxa ($M_l(19)$ < 0.6) in the same way as **FIGURE 3.3**. Additionally, subtrees were arranged such that strains with lower GC % content were placed above, which for the majority of strains could be achieved to fit the general pattern with the exception of *F. perfoetens* (GC % = 26.0). The GC % bounds are 25.2 (*Fa* ATCC 51191) to 35.3 (*Fnec* LS 1195). The GC % content of *I. polytropus* is very similar to *Fnec* at 34.4 %. The horizontal scale represents evolutionary distance based on the average $M_l(19)$ scores between groups.

3.8 Discussion

Using these methods of analysis, it is apparent that the current accepted taxonomic boundaries for *Fusobacterium* spp. must be redefined to better appropriate species definition. The proposed removal of *Fn* subsp. *fusiforme* has existed since its inception where it was incorrectly placed outside of the *vincentii* subspecies where it clearly belongs (19, 20, 191, 192); this was the first of many cases of misnomers within the genus. There is an inherent data bias for the *Fusobacterium* genus that adds to the struggles; there are more data for *Fn* (prior to redefinition) alone than all the remaining strains in *Fusobacterium*. This is likely because most of the strains isolated and sequenced predominantly come from *Fn* and *Fnec*, as these two are the most common human pathogens.

It has previously been stated that *F. naviforme* ATCC 25832 strain had a GC % content higher than other members of the species at 49 % (52.8 % calculated directly from genome assembly), whereas typically it should be 25-35 % for *Fusobacterium*. However, this strain is the only one of the species to have been sequenced, therefore no definitive remarks can be made concerning the remaining strains within the species.

Using the classifications derived in this study, the whole genus of *Fusobacterium*, for which full genomes are available, can be split into at least 20 distinct species (as seen in **FIGURE 3.3**): *animalis*, *closii* sp. nov., *gonidiaformans*, *hwasookii*, *massiliense*, *mortiferum*, *necrophorum*, *nucleatum*, *oralis* sp. nov., *ovarium* sp. nov., *perfoetens*, *periodonticum*, *polymorphum*, *pseudoperiodonticum* sp. nov., *pseudovarium* sp. nov., *russii*, *ulcerans*, *varium*, *vincentii*, and W1481. Henceforth, *Fna*, *Fnn*, *Fnp*, *Fnv* and *Fnw* will be referred to as *F. animalis* (*Fa*), *F. nucleatum* (*Fn*), *F. polymorphum* (*Fpoly*), *F. vincentii* (*Fv*) and *F.* W1481 (*Fw*) respectively. The most notable findings were the identification of two novel species within our strain collection, *F. oralis* sp. *nov*. and *F. ovarium* sp. *nov*., the former of which only had one comparative unclassified genome in the database prior to this study and the latter having no sequenced homologous strains at all. Additionally, various

misnomers were addressed to conform to the newly set species boundaries, such as the removal of the *F. equinum* species, which was internalised by *F. gonidiaformans*.

The reclassifications made in this research closely mirror other studies that have looked at *Fusobacterium* and all bacterial classifications (22, 23). The combination of new data and the examination of the genus as a whole, provide a much more in-depth picture with respect to *Fusobacterium* compared to the other studies, which focussed on just *F. nucleatum* or every bacterial species. An example of this can be seen when examining a phylogenetic tree representing every strain (at the time of writing) that cluster around the previous *F. nucleatum* species (**FIGURE 3.8**). Here, we can see the wide diversity within a group of bacteria once considered its own species (excluding *Fh* and *Fperio*).

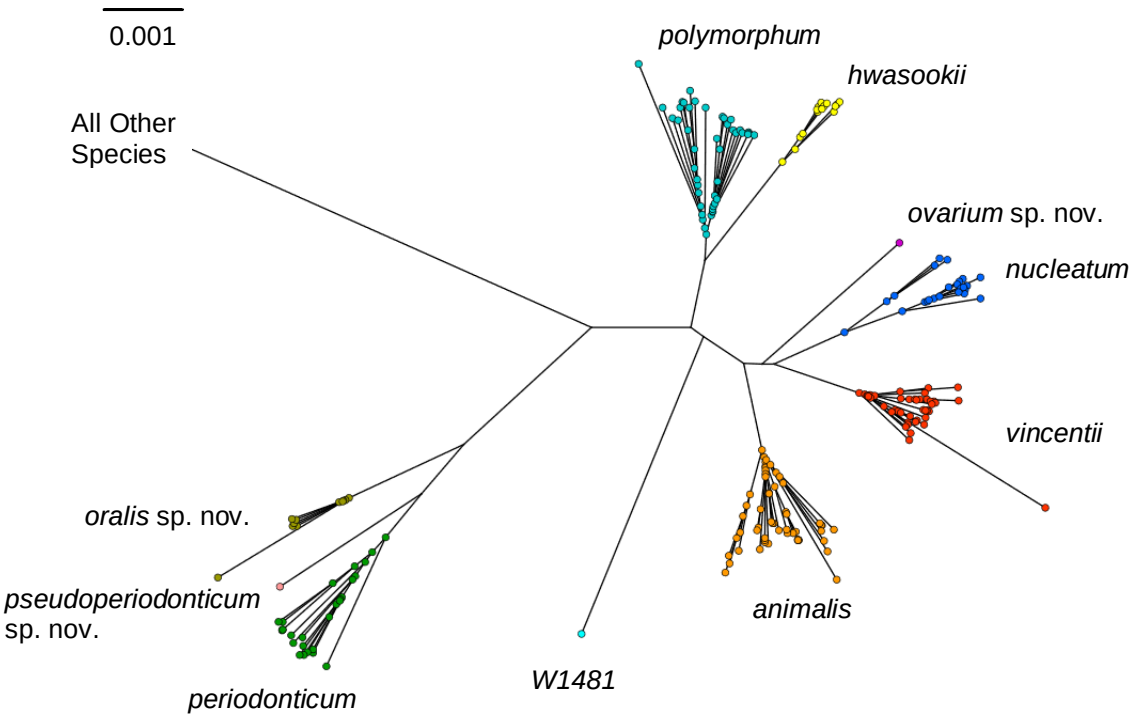


Figure 3.8 | ***F. nucleatum* sub-genera-related species phylogenetic tree.**

This cropped tree displays the species spread about the *Fusobacterium* that cluster around *F. nucleatum*. Each node is coloured uniquely to the related species. The branch lengths correspond to the relative distance as calculated with the MUMi linear transform metric (M_l). The unrooted tree was calculated and drawn using the Neighbour-Joining method in SplitsTree4.

3.8.1 The Duality of *F. polymorphum*

When examining the organisation of *F. polymorphum* species in **FIGURE 3.8**, there is a clear split with two dominant clades emerging from this. Interestingly, there are a few comparisons from the far edges of these two groups where the pairwise score crosses the species threshold. This discrepancy is not as severe as it would be as there are 24 strains within the species – this increases power when judging classifications and means that it can remain a single species if an adequate type strain to compare against is used. As there are multiple candidate strains that do not leave the species definition when compared against any of the existing strains, such a type strain can exist, therefore keeping the species intact.

This emergent duality, however, should not go unnoticed as it suggests an evolutionary split that caused the bacteria to diversify into one of two types. With more data and careful curation, the difference in the two linages could be identified, such as a change in pathogenicity or adaptations to different environments.

3.8.2 A New *F. nucleatum* Subgroup

In **FIGURE 3.8** there is a small subgroup that can be seen within *F. nucleatum*, which is juxtaposed to the main cluster of strains. Previously, this group only consisted of one strain, however, with the new sequencing data, is joined by two other strains. Similarly to *F. polymorphum*, this highlights a divergent evolutionary step that occurred, giving rise to this distinct subgroup. There is no correlation between isolation site and any three of these strains, with one coming from periodontitis, one from a liver abscess and the last from an arthritic hip. As there are few strains within this group, not much can be concluded regarding disease association.

3.8.3 Reorganisation of the *Fusobacteriaceae* family

The overall organisation of the *Fusobacterium* genus shows that each species belongs to a higher order group within the genus. This higher order clustering could contribute partly to the reason why the *nucleatum* species originally contained several subspecies, as the

strains clustered on a cladogram with a distinct separation from *necrophorum*. Though on reflection, this was incorrect; the entire clade of *necrophorum* was spread over a smaller distance than the subspecies of *animalis* for example. The three higher order groups that emerge within the genus are: *necrophorum* and *gonidiaformans*; *animalis*, *nucleatum*, *polymorphum*, *vincentii*, *hwasookii*, W1481, *periodonticum*, *oralis* sp. nov. and *ovarium* sp. nov.; *ulcerans*, *varium* and *mortiferum*. *F. perfoetens*, *F. massiliense*, *F. russii* do not belong to any of these groups and are alone in their respective clades and more data is needed to confirm their positions within the genus.

Analysis comparing *Cetobacterium*, another genus within the *Fusobacteriaceae* family, demonstrated equivalencies to the most distant strains used in the original analysis. For example, *C.* sp. ZOR0034 and *F. ulcerans* ATCC 49185 $A_{abs} = 72.25\ \%$, which is about 2.4 % higher than the lowest score seen in the *Fusobacterium* data (*F. ulcerans* 12-1B and *F. necrophorum* subsp. *funduliforme* F1309 $A_{abs} = 69.83\ \%$) and is higher than 191 other intra-genus comparisons. This discontinuity is also evident in **FIGURE 3.7** where *F. necrophorum* clusters outside of the previous genus boundary, existing in a higher order above that of *Cetobacterium*.

Based on the findings presented in this study as well as the observations made in the global bacterial taxonomic study (23), *Fnec*, *Fg, Fvar, Fpv*, *Fu*, *Fmort*, *Fperf,* and *Fc*, should all be removed from the *Fusobacterium* genus and reclassified into at least two separate distinct genera. This would lead to three genera within the *Fusobacteriaceae* family previously all grouped under one. Based on the clustering observed in **FIGURE 3.7**, these new genera would contain the following species: (1) *Fn*, *Fa*, *Fv*, *Fpoly*, *Fw*, *Fov*, *For*, *Fh*, *Fperio*, *Fpperio*, *Fmal*, and *Fr*; (2) *Fu*, *Fvar*, *Fpv*, *Fmort*, *Fperf*, and *Fc*; (3) *Fnec* and *Fg*.

Chapter 4: *Fusobacterium* TAAs and Adhesion to CEACAMs

4.1 Introduction

As previously explained in **Section 1.3**, CEACAM1 (Carcinoembryonic Antigen Cell Adhesion Molecule 1) can make an ideal pathogen-binding receptor with the potential to confer local immune suppression. From prior unpublished studies conducted by this research group, it was found certain *Fusobacterium* strains could bind this receptor; these interactions were limited to *F. nucleatum* and *F. vincentii* strains as well as some previously uncharacterised clinical isolates. In *Fn* ATCC 25586, a protein, encoded from the gene locus FN1499, was found to be responsible for mediating the interaction with CEACAM1. This protein was identified by coprecipitation of bacterial lysates with CEACAM1, followed by N-terminal sequencing using Edman degradation and mass spectroscopy of the corresponding protein bands on an SDS-PAGE gel.

The protein product of FN1499 was predicted to belong to the Trimeric Autotransporter Adhesin (TAA; also known as Type Vc Secretion System) family of proteins as described in **Section 1.2**. This prediction was made by identifying a classical T5SS 12-stranded β-barrel in the C-terminus of the predicted coding sequence. This is further backed up by the presence of YadA-like head domains and the presence of a neck domain, however, the stalk region of this protein remains undefined and there are no close homologues for it in other non-*Fusobacterium* strains when examining BLAST results.

TAAs can be found within all members of the *Fusobacterium* genus, though their roles have yet to be characterised. *Fn* ATCC 25586 contains three gene loci that encode putative TAAs, FN0471, FN0735 and FN1499, the latter of which has been shown to bind human CEACAM1. While TAAs are ubiquitous among the genus, homology to these three protein sequences is varied. For example, a homologue to FN0735 can be found in *Fn*, *Fv*, *Fa*, and *Fpoly*, previously all classified under the umbrella of *F. nucleatum*. FN0471 homologues

appear to much less prevalent, with only *Fn* and *Fv* strains showing possession. In addition to the close homologues originating from *Fn* for example, more distant species such as *F. necrophorum* also possess genes encoding putative TAAs, which are far-removed from proteins encoded in *Fn* for example. As explored in **SECTION 1.2.3**, TAAs have been shown to be important in bacterial adhesion to many human cellular components, such as cell-surface receptors or extracellular matrix components, and often play an important role in pathogenesis. Successful characterisation of these proteins may indicate whether they are suitable and targetable as vaccination candidates.

Functional homologues to FN1499 were identified in *F. vincentii* and in several of the clinical isolates, such as strain 2B3, since classified as *F. oralis* sp. *nov.*, that were also able to bind CEACAM1. The 2B3 strain exhibited a slightly longer protein of 519 amino acids compared to 479 residues in FN1499 with molecular weights of mature (lacking signal peptide) protein monomers at 51.2 and 48.1 kDa respectively. This contrasts to the homologues found in *Fv* strains which bared a much closer resemblance to FN1499 from *Fn*. These proteins will henceforth be denoted by the term CbpF for CEACAM-binding proteins of *Fusobacterium* – FN1499-like proteins as CbpFa and 2B3-like as CbpFb.

The overall predicted topology of CbpFs, as well as other TAAs within *Fusobacterium* spp. is relatively simple when compared to TAAs from other species such as *S. enterica* or *M. catarrhalis* where the respective proteins SadA and UspA1 are large complex multidomain proteins (102, 108). The predicted architecture, which is investigate in depth in **SECTION 5.2**, contains only the core elements required for a TAA: a membrane anchor, a single stalk followed by a neck and then a series of head domains. In the case of FN0735 and its homologues, there is additional stalk and neck domain. All of the head domains adopt the YadA-like motif and no other more exotic features, listed in **TABLE 1.1**, were identified; however, it is possible that *Fusobacterium*-derived TAAs have novel structures yet to be characterised. A schematic topology of CbpFs is displayed in **FIGURE 4.1**, showing the

domain organisation as predicted using sequence analysis described in detail in **SECTION 5.2**.

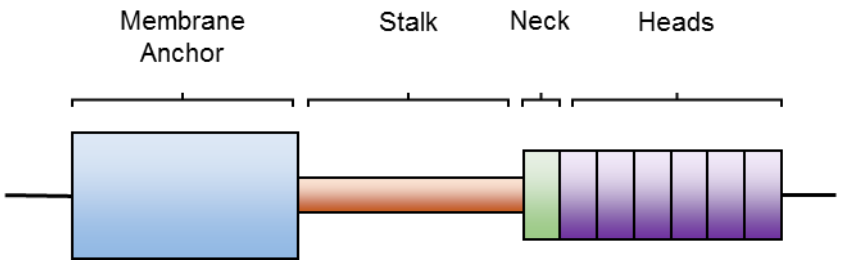


Figure 4.1 | **CbpF domain topology.**

CbpFs are trimeric proteins that contain four core domains identified in sequence analysis: a membrane anchor consisting of a 12-strand β-barrel, a stalk region (likely a coiled coil), a neck, and series of head domains all adopting a YadA-like head motif. These domains are the minimal requirements for any TAA and in this case, show one of the simplest architectures.

Interestingly, CbpFb has no direct homologues within the *Fn* or *Fv* species – in fact the closest match in the current database (excluding newly characterised strains within this study) belongs to that of an *F. animalis* species (*Fa* CAG:649). This agrees with our previous findings in **CHAPTER 3** that the 2B3 strain belongs to a distinct species within *Fusobacterium* – *F. oralis* sp. nov. – which is genetically further from *Fn* and *Fv* by approximately the same distance as *F. periodonticum*. In contrast, a CbpFa homologue appears in all *Fn* and *Fv* species, with the exception of where the gene lies on the end of an incomplete contig, such as in the case of *Fn* CTI-2, where a partial hit is found spanning two contigs. It is likely that this gene is present within all other non-sequenced *Fn* and *Fv* strains. This could therefore have implications in disease, as it appears to be completely conserved among these species and could have a role in host specificity or even be directly involved in pathogenesis, both of which will be discussed later.

In addition to CEACAM1 binding, there is little data on whether these CbpF proteins can adhere to other members of the CEA family, which could include other CEACAMs, such as 3, 5 or 6, which are known to bind to other pathogens, such as *M. catarrhalis* (118, 134).

Alternatively, there could be interactions with pregnancy-specific glycoproteins (PSGs), which form a large understudied branch of the CEA family of receptors. Any interactions with these additional receptors could have strong implications with pathogenesis and disease specificities.

In this chapter, we aim to characterise the binding profiles of CbpFa and CbpFb with different CEACAMs. In addition, we will explore the CbpF-CEACAM1 binding interface at cellular and molecular levels using a variety of methods which will be described in turn.

4.2 CbpF-CEACAM Interactions

4.2.1 CEACAM1 Screening in *Fusobacterium* Clinical Strains

To confirm CEACAM1-binding of strains that possess a predicted CbpF homologue within their genome, a combination of data from previous unpublished studies and data gathered in this study were used to cross-validate results. Data were collected by detecting adhesion between a soluble CEACAM1 conjugate protein with bacterial cell lysates using Western blots. The 25 clinical strains that were used for WGS and some type-strains of other *Fusobacterium* species were used in these experiments. The control type strains used were: *Fn* ATCC 25586, *Fv* ATCC 49256, *Fa* ATCC 51191, *Fpoly* ATCC 10953 and *F. pseudoperiodonticum* sp. nov. ATCC 33693. From which, only the *Fn* and *Fv* type-strains displayed CEACAM1 adhesion.

The protocol used for lysate preparation and blotting are described thoroughly in the Methods. *Fusobacterium* lysates were prepared and were then run on a 4-20% SDS-PAGE gel before either staining with Coomassie or transferring to a nitrocellulose membrane. Western blots were then carried out, however the primary antibody used was, instead, the CEACAM1-4C1-hIgG1 F_C conjugate (CC1-F_C) which was detected with an anti-human IgG-AP secondary antibody. In addition to this, a CC1-F_C mutant lacking the N-terminal IgV domain of CEACAM1 (CC1 ΔN-F_C) was used as a negative control, as the IgV-like domain of CEACAM1 is very likely responsible for adhesion. An example of the blots is shown in

FIGURE 4.2 comparing CbpFa and CbpFb-harbouring strains. **TABLE 4.1** shows the full consensus of results from this study as well as previous work conducted by Hill *et al* (unpublished), from the clinical isolates and their CEACAM1-binding capability.

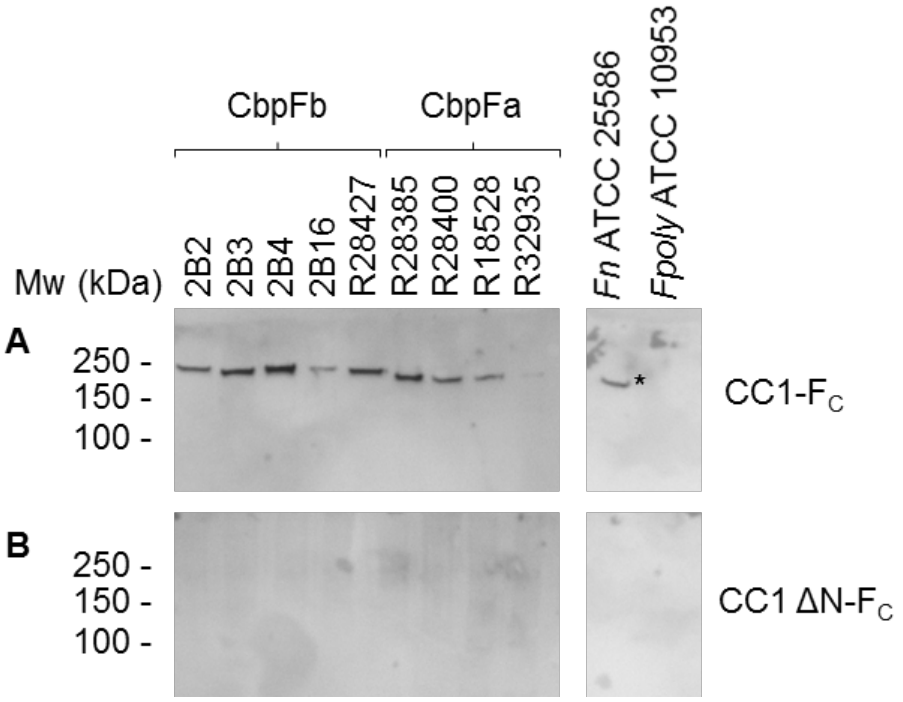


Figure 4.2 | ***F. oralis* sp. nov. CEACAM1-binding profile compared to *F. nucleatum*.**

Displayed are Western blots of *Fusobacterium* lysates from 5 *F. oralis* and 4 *F. nucleatum* clinical strains overlaid with CC1-F_C (**A**) or CC1 ΔN-F_C (**B**) and anti-human IgG-AP. Included as controls were *F. nucleatum* ATCC 25586 (***** – CDS product of FN1499, predicted 144.3 kDa) and *F. polymorphum* ATCC 10953 (harbours no predicted binding protein). As can be seen, the *F. oralis* CEACAM1-binding proteins are slightly larger than the *F. nucleatum* proteins, which is to be expected as a CbpFa trimer is approximately 10 kDa smaller than a CbpFb trimer (144 and 154 kDa respectively). *F. oralis*: 2B3, 2B4, 2B16 and R28427. *F. nucleatum*: R28385, R28400, R18528 and R32935. CC1-F_C – CEACAM1-4C1-hIgG1 F_C; CC1 ΔN-F_C – CEACAM1 A1BA2-hIgG1 F_C; AP – Alkaline phosphatase. A total of three blots were carried out in this study – shown is an example of one.

Table 4.1 | **Clinical strain CEACAM1-binding profiles.**

The CEACAM1-binding characteristics of all the clinical isolates used in the whole-genome sequencing study was determined across several studies with confirmation of some of the strains in this study by Western blot. [1] As determined in **CHAPTER 3**. [2] Consensus of results from data gathered in this study and unpublished work. [3] Variable binding observed. [4] Likely a mixed culture of *F. vincentii* and a *Bacillus* species.

Strain	**Species[1]**	**CEACAM1 Binding[2]**
2B16	*F. oralis*	+
2B17	*F. vincentii*	+
2B2	*F. oralis*	+
2B3	*F. oralis*	+
2B4	*F. oralis*	+
R15792	*F. animalis*	+/- [3]
R16531	*F. ovarium*	+
R18528	*F. nucleatum*	+
R18932	*F. animalis*	-
R24394	*F. nucleatum*	+
R26872	*F. vincentii*	+
R28211	*F. vincentii*	+
R28385	*F. nucleatum*	+
R28400	*F. nucleatum*	+
R28427	*F. oralis*	+
R29976	*F. vincentii*	+
R30464	*F. vincentii*	+
R30604	*F. vincentii*	+
R30927	*F. animalis*	-
R31249	*F. vincentii*	+
R32310	*F. vincentii*	+
R32935	*F. nucleatum*	+
R33458	*F. vincentii/Bacillus* sp.[4]	+
R33533	*F. vincentii*	+
R5001	*F. animalis*	+

Of the strains tested, all the strains identified as *F. nucleatum* or *F. vincentii* bound to CEACAM1, although some variation was observed. Variable binding could be due to a variety of factors such as uncharacterised phase or antigenic variation, protease action, or total protein differences caused by naturally low or high expression of CbpF. In **FIGURE 4.2**, there are clear bands that correspond to the predicted trimeric forms of CbpFa and CbpFb (144 and 154 kDa respectively), however, the bands are approximately 100 kDa larger than the predicted molecular weight. This is because these proteins cannot be dissociated and denatured using SDS-PAGE loading buffer alone (with 10 min at 95 °C), as the stability of the membrane anchor is very high and harsher treatments, such as heating in the presence of formic acid, are required to break apart this domain. Monomeric protein units would appear in the 50 kDa region.

All the *F. oralis* strains, 2B2, 2B3, 2B4, 2B16 and R28427, were able to bind CEACAM1, which confirms CEACAM1-binding is not limited to *Fn* or *Fv* species within *Fusobacterium*. Moreover, there were also two *F. animalis* strains that could bind CEACAM1; these were R15792 (though variably) and R5001 (**FIGURE 4.3**). These two particular strains harboured a CbpF that resembled that of CbpFb but were distinctly different. This indicates that a third group of CEACAM1-binding proteins could exist. Interestingly, however, the sequence of R5001 is much closer to that of CbpFb than that of R15792 which contains regions that closer match CbpFa, indicating they may have obtained the gene at different points in their evolution.

FIGURE 4.4 shows the phylogram representing the species of *F. animalis* and where the identified CEACAM1-binding strains fit into the underlying taxonomy. It is immediately evident that all the strains able to bind CEACAM1 are in a separate clade compared to the main group of the species, suggesting these strains have an inherent difference as well as being the only identified *animalis* strains able to bind CEACAM1. Interestingly, not all the other strains within this subgroup can bind CEACAM1 with R30927 not being able to physically bind CEACAM1 as well as lacking any discernible homologue within the

sequence. In addition to this, the previously unclassified strain, *Fusobacterium* CAG:649 also falls into this subgroup of *F. animalis* and a CbpF-like gene can be identified within the sequence.

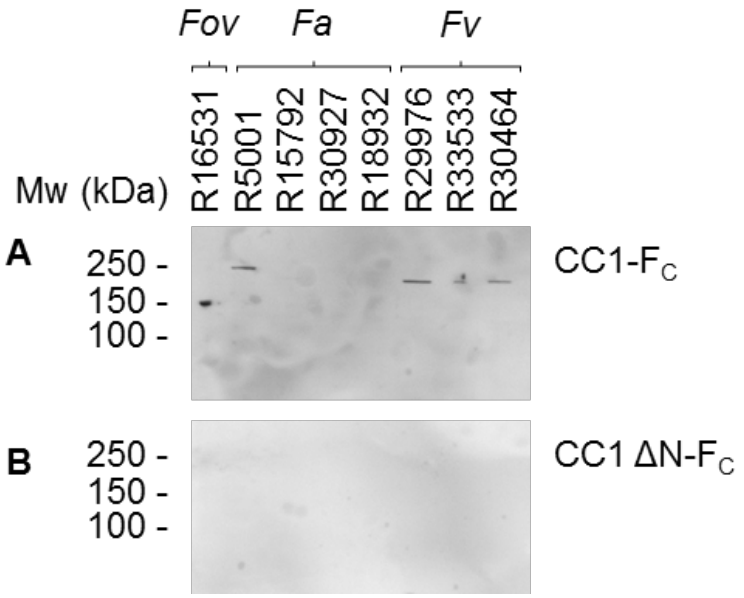


Figure 4.3 | ***F. ovarium* sp. nov., *F. animalis* and *F. vincentii* CEACAM1-binding profiles.** Western blots of lysates from *F. ovarium* sp. nov. (*Fov*), *F. animalis* (*Fa*) and *F. vincentii* (*Fv*) strains against (**A**) CEACAM1-F_C (CC1-F_C) and the (**B**) CEACAM1 N-terminal IgV deletion mutant (CC1 ΔN-F_C). The *Fv* CEACAM1-binding protein is also from the CbpFa line of proteins, whereas the CbpF proteins in *Fov* R16531 and *Fa* R5001 differ in size with the protein from *Fov* appearing smaller and *Fa* larger than CbpFa. Blots were repeated three times; however, the displayed blot shows R15792 being unable to bind, which conflicts with previous findings.

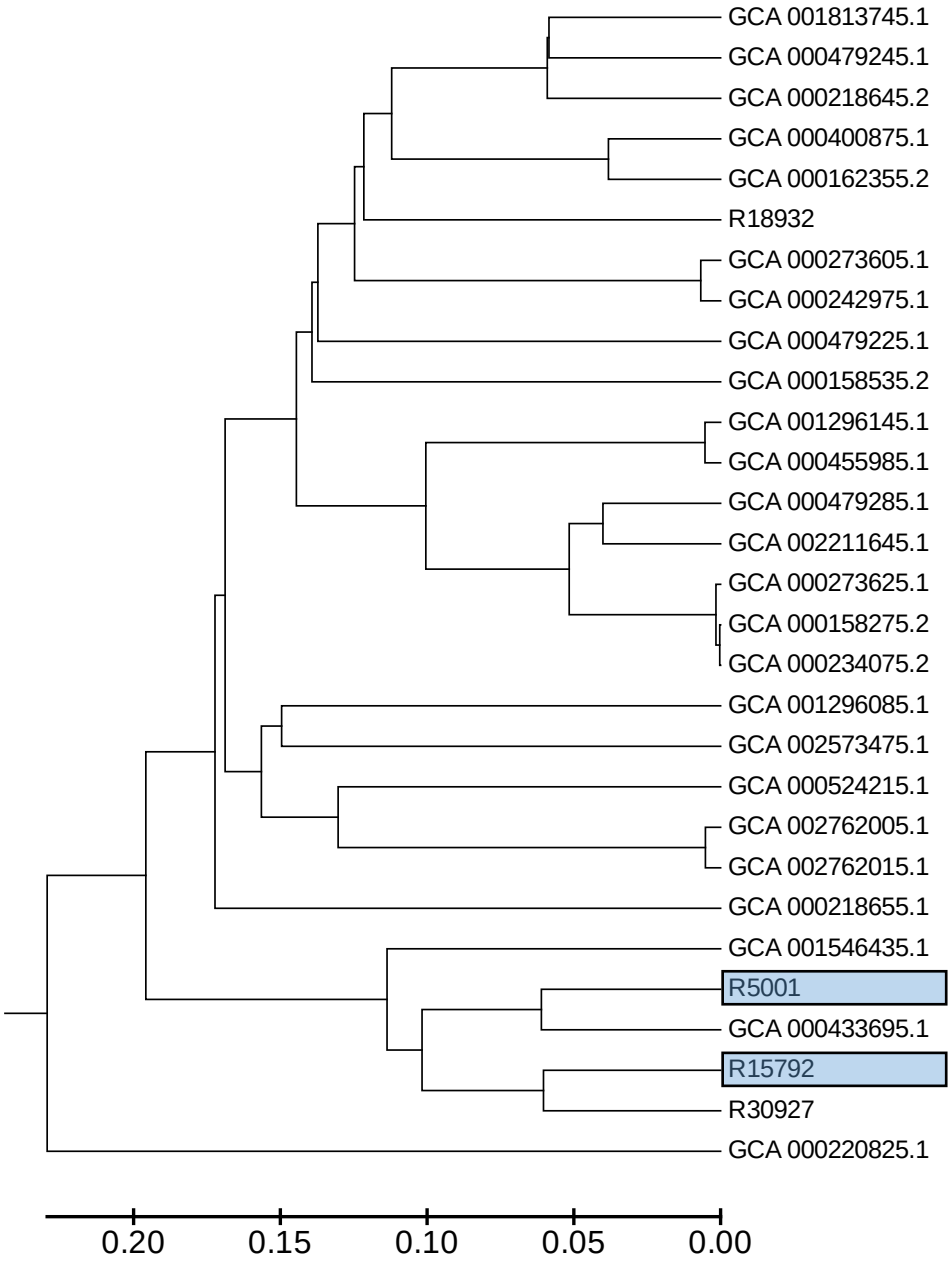


Figure 4.4 | **The *F. animalis* species distribution.**

The phylogenetic tree represents whole-genome comparisons using MUMi restricted to *F. animalis* strains. Highlighted in blue boxes are the two confirmed CEACAM1-binding strains. GCA_000433695 and GCA_001546435 (*Fusobacterium* sp. CAG:649 and *Fusobacterium* strain MJR7757B respectively) also have a CbpF homologue and can likely adhere to CEACAM1 too. The R30927 clinical strain does not contain a CbpF homologue in its genome as well as not being able to bind CEACAM1. The GenBank assembly accession numbers are given for the genomes used. The UPGMA method was used to infer evolutionary distance from the linearized MUMi transformation metric (LMUMi) between strain pairs with the tree generated using MEGA7.

The final species confirmed to bind CEACAM1 in this study was *F. ovarium* sp. nov. (*Fov*). This strain is interesting as it exists as a singleton set apart from any other species though its closest relatives are *Fn*, *Fv* and *Fa*. In addition, there are more than one predicted CbpF-like genes in this strain, each of which also has some unique properties not found in the other CbpFs. The first of these proteins is heavily truncated in the predicted stalk region and has only 387 amino acids, approximately 100 residues shorter than CbpFa. The second CbpF-like protein is more similar to CbpFa, but with two insertions with a resulting total length of 488 residues. These two proteins will be referred to as putative CbpFc1 and CbpFc2 respectively until CEACAM-binding profiles can be established for each individual protein.

FIGURE 4.3 shows the results of a Western blot comparing the *Fov* and *Fa* CEACAM1-binding strains, as well as a comparison to *F. vincentii*. In this case, the *Fa* R15792 strain failed to demonstrate visible adhesion, while the *Fa* R5001 and *Fov* R16531 strains both showed CEACAM1-specific interactions. This experiment also shows that the size of the CEACAM1-bining protein in *Fov* is markedly smaller than that of CbpFa and CbpFb, indicating that the CbpFc1 (with a trimer size of approximately 115 kDa) is the likely candidate for CEACAM1 binding.

The one common feature all the *cbpF* genes have (barring *cbpFc2*), is their proximity to certain surrounding genes. The c*bpF* gene is always found directly upstream of the *nik*-operon (*nikABCDE*) followed by *ribD* and *ribH*. In addition to *cbpF* being restricted to CEACAM1-binding strains, these strains are also the only ones harbouring an intact *nik*-operon. This is indicative of this whole region, at one point in time, being a mobile genetic element and evidence of a transposase further downstream of these genes can also be seen in some strains. The exception to this is the putative *cbpFc2* gene in *F. ovarium* sp. nov. where this gene is inserted into a region downstream of a histidine kinase gene.

FIGURE 4.5 shows the genetic map of the different CbpFs and how their genomic topology compares. As can be seen, not all elements upstream of the *cbpF* gene are present, though the downstream *nik*-operon is more or less the same in all strains.

Contrary to the presence of the *nik*-operon, there is no evidence of the *nikR* gene present in the genomes of any *Fusobacterium* strain. The gene product NikR is responsible for regulation of the *nik*-operon (product NikABCDE) and is found within other strains that harbour this operon, which is responsible for nickel transport using an ABC-transporter. However, it is clear that this particular operon is not needed for successful growth or proliferation of *Fusobacterium* as it is not found in the vast majority of strains. It is unclear what the relationship between this operon and CbpF is, where it may be an incidental relationship that has no effect on one another or that the expression of one will lead to increased expression of the other for example.

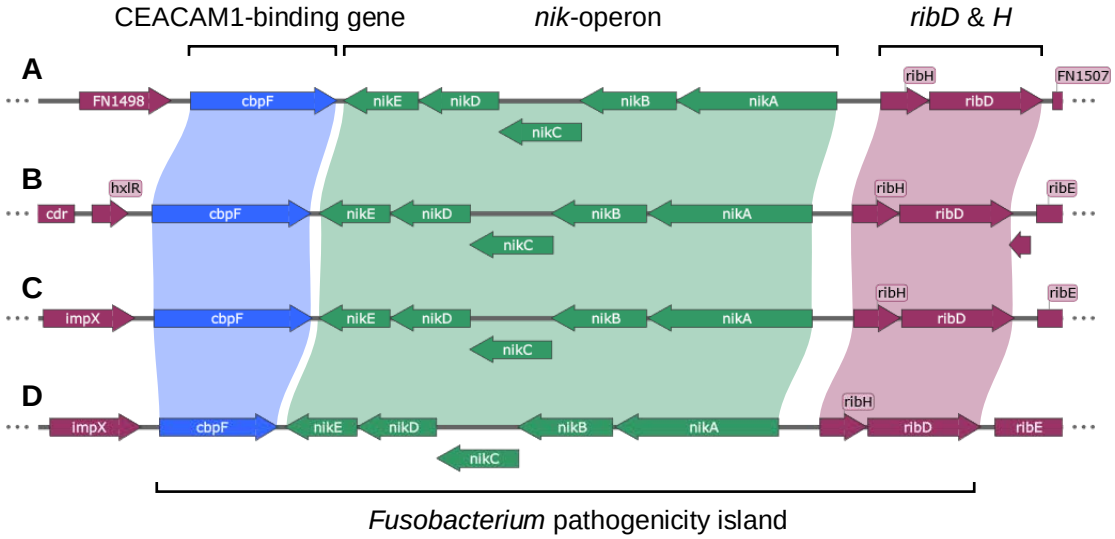


Figure 4.5 | **The CbpF genomic island.**

The above map shows the position of the *cbpF* gene within **A)** *F. nucleatum* ATCC 25586, **B)** *F. oralis* sp. nov. 2B3, **C)** *F. animalis* R5001 and **D)** *F. ovarium* sp. nov. R16531 genomes. The *cbpF* region shown for *F. ovarium* sp. nov., is the area surrounding the *cbpFc1* gene. The *cbpF* gene is found directly upstream of the *nik*-operon in each case with the *ribD*, *H* and *E* genes further downstream of this. The regions of DNA represent 10000 bp each and are shown to scale.

FIGURE 4.6 shows the evolutionary history of identified putative CbpF proteins within *Fusobacterium*. Here the two predominant branches of CbpF proteins can be seen as well as the internal variation within the CbpF types. From examining the sequence alignment (see **APPENDIX D**) of all the predicted CbpF proteins, it is clear one domain is completely conserved, that being the β-barrel coding region. The sequence then differs throughout the remaining domains in the protein with the primary difference being altering numbers of YadA-like domains in the head portion of the protein. The signal peptide is the other highly conserved region, with few changes in sequence, though this is to be expected for successful trafficking to occur. The stalk region also differs between strains, though it hard to make any predictions about how this would affect function, as this domain cannot be accurately characterised.

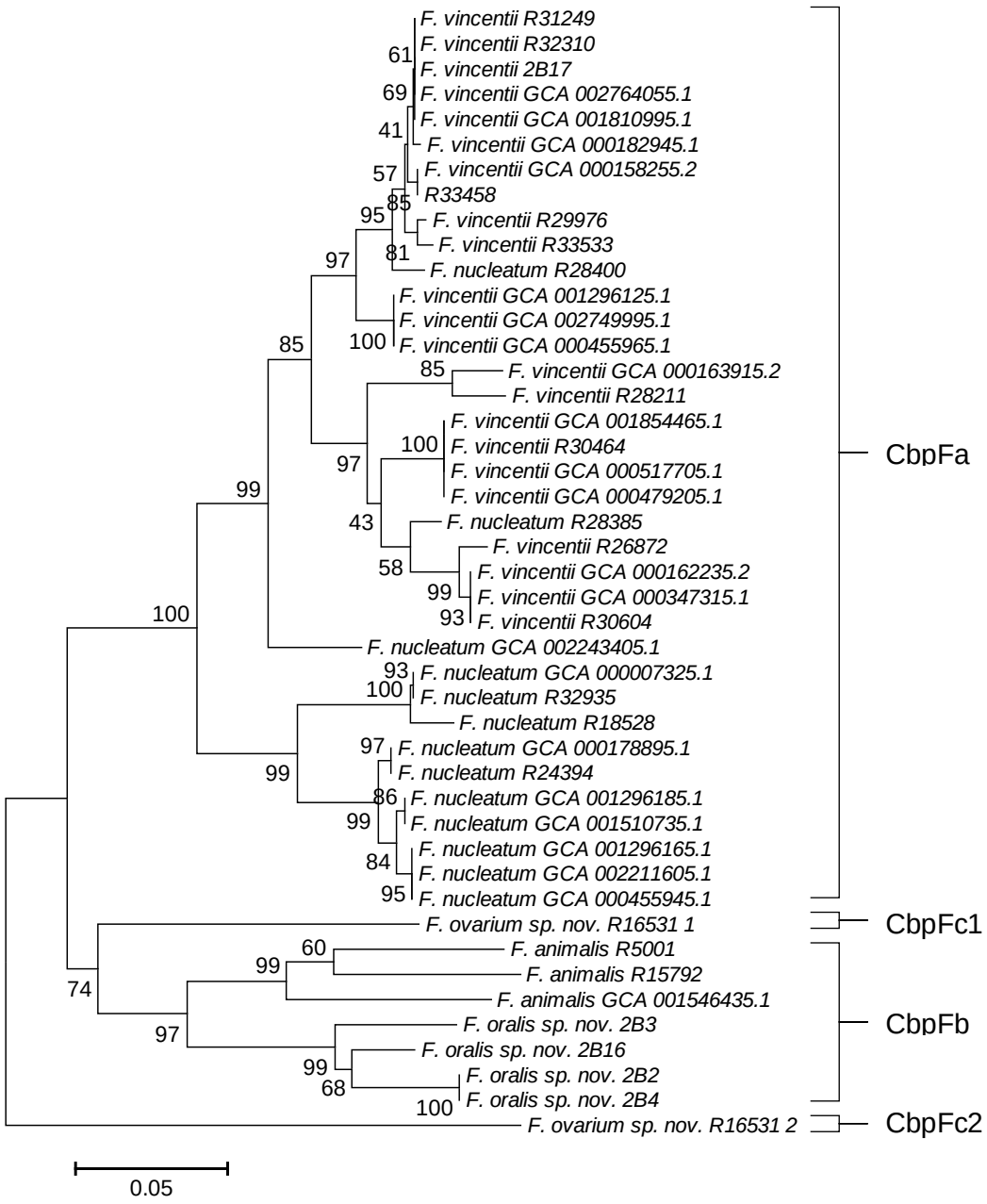


Figure 4.6 | **CbpF Evolutionary History.**

Putative CbpF proteins were identified within the above strains and aligned using MUSCLE (166). The evolutionary history was inferred using the BioNJ method in MEGA7 (167) and the optimal tree, with the sum of the branch lengths equal to 1.194, is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (193), with 1000 replicates, are shown next to the branches. The evolutionary distances (number of amino acid substitutions per site) were computed using the Poisson correction method as used in MEGA7 (167). Strains with partial CbpFs due to existing on contig ends, such as *Fn* CTI-2 and *Fa* CAG:649 were excluded.

4.2.2 Enzyme-linked Immunosorbent Assay with Purified Protein

To examine the effects of CEACAM interactions more closely, assays were performed using purified protein. Both CbpFa and CbpFb from *Fn* ATCC 25586 and *For* 2B3 respectively were cloned in to pOPINE (pCRF1 and pCFR2 respectively) as described in the Methods and subsequently transformed into *E. coli* expression cells. Initially, BL21 (DE3) cells were used for expression and protein purified using the small-scale method. This, however, yielded very little protein, so the cell line was switched to BL21 (DE3) pLysS cells, which contained a plasmid responsible for preventing 'leaky' expression prior to induction. The induction time was also changed to O/N at RT. Protein was then purified using the same method as before and this greatly improved protein yield. Purified protein was subsequently exhaustively dialysed against PBS and kept at a concentration below 0.1 $mg \cdot ml^{-1}$ and at RT to prevent aggregation and precipitation. CEACAM1 binding activity was confirmed using a Western blot using CEACAM1-F_C (CC1-F_C) in addition to using CbpF-specific polyclonal rabbit serum (PA5154). **FIGURE 4.7** shows the activity and purity of purified recombinant CbpFa 22-330.

ELISAs were then performed using the purified proteins against CEACAMs where the total amount of protein used to coat the wells was 3 pmol of each. The primary detection was done using soluble human CEACAM1, CEACAM3, CEA, CEACAM6, CEACAM8, CEACAM1-ΔN and mouse CEACAM1b hIgG1-F_C conjugates at a concentration of 1 pmol per well as well as a blank 1 % (w/v) BSA negative control. Binding of these conjugates was detected using anti-human IgG alkaline phosphatase (āH-AP) antibody followed by development with SigmaFast® (see Materials and Methods). Absorbance at 405 nm was read after 45 minutes of development at 37 °C and the results standardised for each repeat (**FIGURE 4.8**).

The experiment was repeated three times and ANOVA followed by a post-hoc Tukey HSD test was performed on the results to identify the significant differences. The graph of results in **FIGURE 4.8** confirms binding of CbpFa and CbpFb to CEACAM1, but also shows that both

are able to bind CEA, however no significant detection cold be observed using any of the other CEACAMs used in this experiment. Furthermore, it shows CbpFb binds to CEA significantly ($p < 0.0001$) more strongly than CbpFa with about twice the absorbance observed, and both CbpFs bind CEACAM1 significantly more than CEA ($p < 0.0001$), though to a lesser extent for CbpFb.

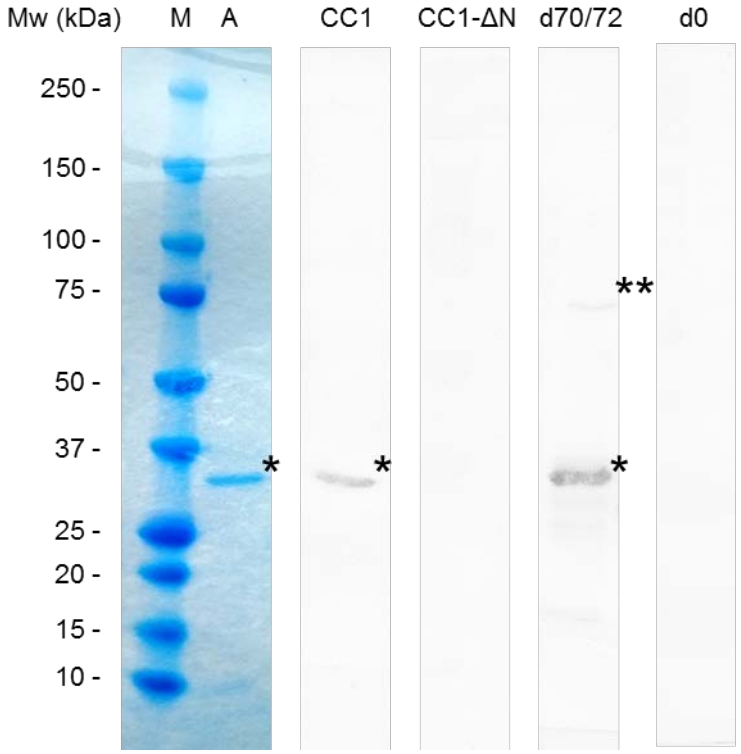


Figure 4.7 | **Conformation of purity and activity of recombinant of CbpFa.**

Recombinant CbpFa was heated (95 °C) with loading buffer (**TABLE S 1**) for 10 min and run on a 4-20 % gradient SDS-PAGE gel for 30 min at 300 V before staining with Coomassie (left) or transferring to a nitrocellulose membrane and performing Western blots (see Methods) with recombinant CEACAM1-hIgG (CC1), CEACAM1-ΔN-hIgG (CC1-ΔN), polyclonal PA5154 day 70/72 (d70/72), and PA5154 day 0 (d0). * marks the band corresponding to CbpFa 22-330 at 32.0 kDa. ** marks residual dimeric CbpFa at 64 kDa. A faint laddering effect can be seen when examining the polyclonal serum blot indicating possible degradation, however a single distinct band is seen against CC1 indicating any signal from an ELISA will predominantly be from CbpFa. M – marker; A – purified CbpFa.

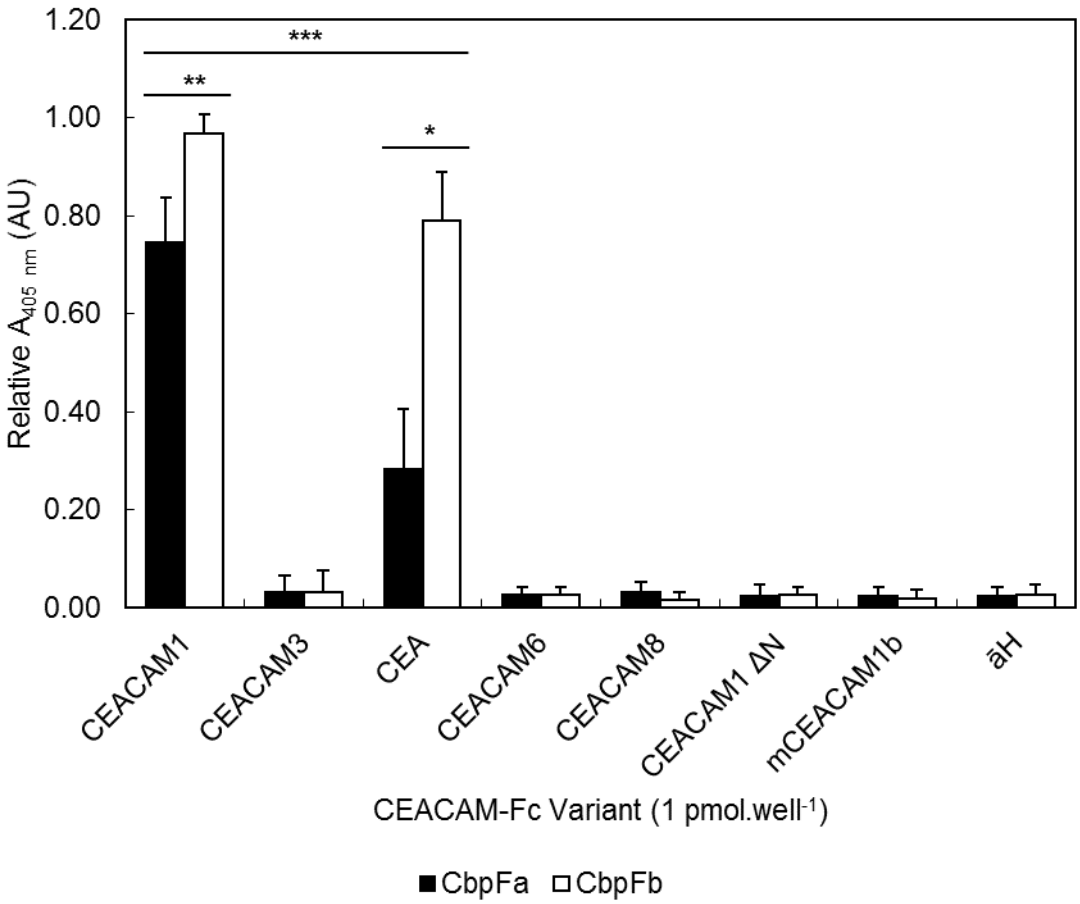


Figure 4.8 | **ELISA examining interactions between CEACAMs and CbpFs.**

3 pmol of purified CbpFa (*Fn* ATCC 25586) or CbpFb (*F. oralis* sp. nov. 2B3) were overlaid on an ELISA plate O/N and blocked with BSA. 1 pmol of each CEACAM-hIgG1 F_C variant were used followed by an anti-human secondary antibody to detect adhesion. Each CEACAM-F_C conjugate contained all the extracellular IgC- and IgV-like domains except for the ΔN mutant that lacked the N-terminal IgV-like domain. A BSA-only control (āH) was included as background negative control. Mouse CEACAM1b (mCEACAM1b) and CEACAM8 were also used as these have not been shown to bind any human pathogen adhesins. Three biological replicates consisting of three technical replicates were performed for each condition. A one-way ANOVA ($F(15) = 370.8;\ p < 0.0001$) followed by a post-hoc Tukey HSD test identified significant differences between CEA and CEACAM1 (******* $p < 0.0001$) binding with CEA binding less for both CbpFa and b. CbpFa was also shown to bind to CEA with about half the affinity compared to CbpFb (***** $p < 0.0001$). A significant difference was also observed between CbpFa and CbpFb binding to CEACAM1 (****** $p < 0.0001$).

4.2.3 Adhesion Assays with CbpF and CEACAMs

In addition to purified proteins, we examined whole-cell binding as well to see if there was any difference when the protein was expressed in its native form on the surface of bacteria. In this study, interactions between CEACAM1, 3, 5 and 8 were examined. To observe these interactions, surface expressed forms of the proteins were used that were engineered to localise to the outer-membrane of *E. coli* using the pOAF expression vector. The full genes (excluding the N-terminal signal sequence) for CbpFa and CbpFb were cloned into this vector using ligation independent cloning as described in the Methods. *E. coli* BL21 (DE3) pLysS cells were transformed with the plasmids as well as empty vector, to assess the background levels of adhesion. These would best represent the *in vivo* forms of the proteins without using *Fusobacterium* as, currently, *Fusobacterium* have no straightforward direct genetic manipulation protocols. We performed an adhesion assay with 100 µl of bacteria per well at a concentration of approximately of 5×10^8 CFU·ml^{-1}, this represents a multiplicity of infection (MOI) of 500. In **FIGURE 4.9**, very good adherence of CbpFa- and CbpFb-expressing bacteria to CEACAM1 was observed, as expected. In addition, we saw some cell-specific interactions of both CbpFa and CbpFb to CEACAM3, though this was not to the same extent as with CEACAM1. For CbpFa and CbpFb, there was also evident cell-specific binding to CEA. These results contrast with the protein-only ELISA data; however, this experiment has more inherent variation due to the much higher number of parameters that could change.

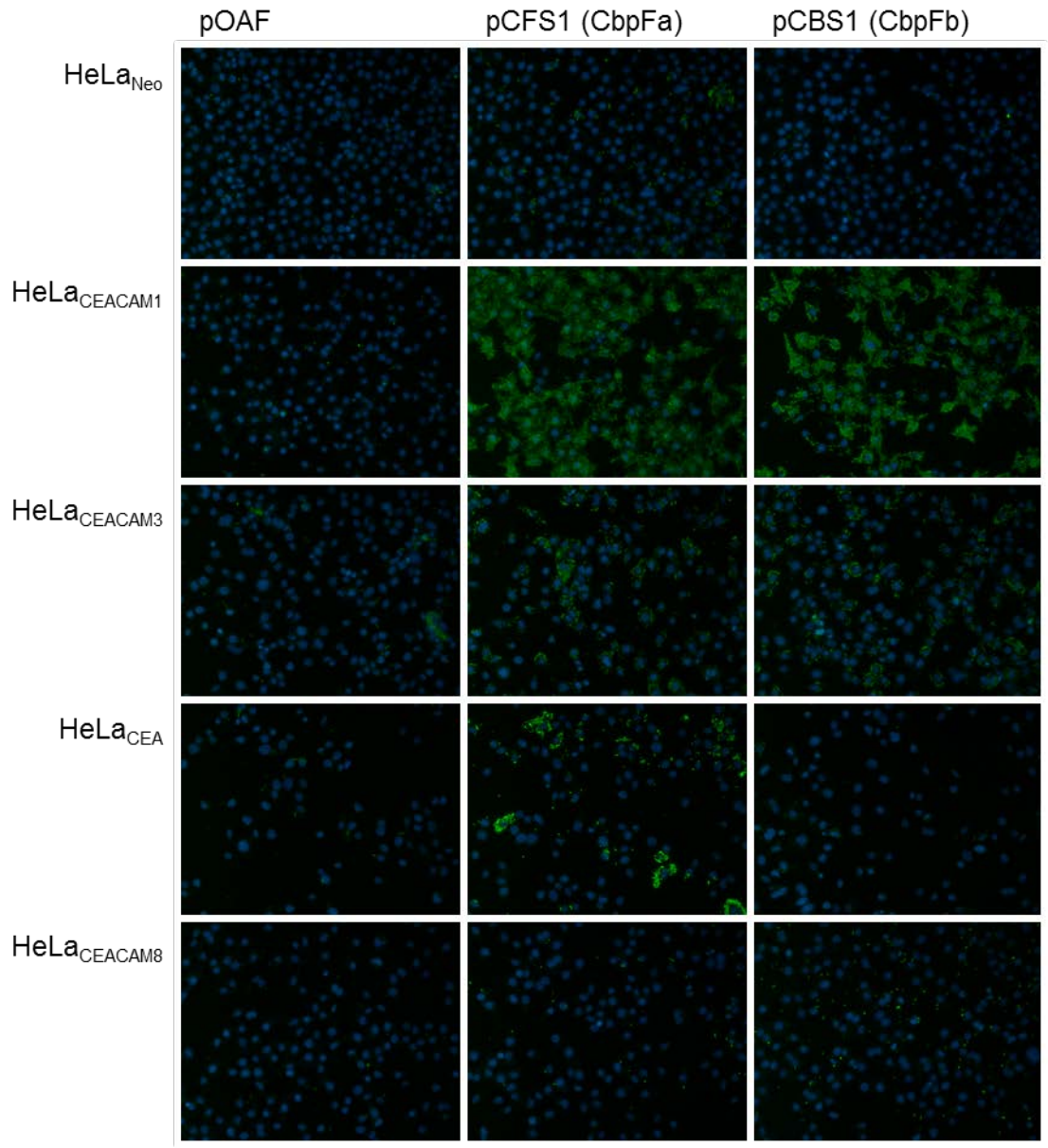


Figure 4.9 | **Whole-cell adhesion assay between CEACAMs and CbpFs.**

An adhesion assay using three *E. coli* BL21 (DE3) pLysS expression strains, harbouring the plasmids pOAF, pCFS1 and pCBS1 (**TABLE 2.3**), and HeLa cells expressing different CEACAMs was performed (see Methods for details). HeLa cell nuclei were stained with DAPI (blue) and bacteria were detected using a fluorescent antibody (green). *E. coli* containing the empty plasmid, pOAF, and non-CEACAM1-expressing HeLa$_{Neo}$ cells were used as negative controls.

4.2.4 Inhibition Study

To study the binding of CbpFa to CEACAM1, we performed an adhesion assay as previously explained, but prior to incubation with bacteria, different potential protein inhibitors were added (100 µl at 1 µg·ml^{-1}) to the cells and incubated for 30 min at 37°C, before being washed off with Medium-199. A purified peptide (rD-7) derived from UspA1 from *M. catarrhalis*, which has been shown to bind to CEACAM1 was used as a potential inhibitor (118). If inhibition was seen with this peptide, then deductions may have been able to be made with respect to where the likely CbpF epitope is located on CEACAM1, as this has been determined for rD-7. In addition to this a media-only condition was setup as well as using another peptide fragment (MsfA) from the *N. meningitidis* protein Msf was used as a protein negative control, as this is a TAA that has not been shown to bind CEACAM1.

FIGURE 4.10 shows the results of the adhesion assay. The media-only control behaves as expected with strong bacterial adhesion displayed and this is also the same for the condition containing MsfA. When looking at the effects of rD-7 on adhesion, there is a marked reduction on the level of adhesion down to basal levels. A one-way ANOVA followed by a post-hoc Tukey HSD test yielded significant differences between rD-7-inhibitded CEACAM1 conditions compared to both Msf-containing and media-only conditions ($p < 0.0001$). Moreover, there was no significant difference between all conditions that remained largely negative or either of the two positive conditions. The experiment collected data from three biological repeats each having three technical repeats per condition. Image data was converted to numerical data using relative fluorescence as a metric for adhesion (see Methods).

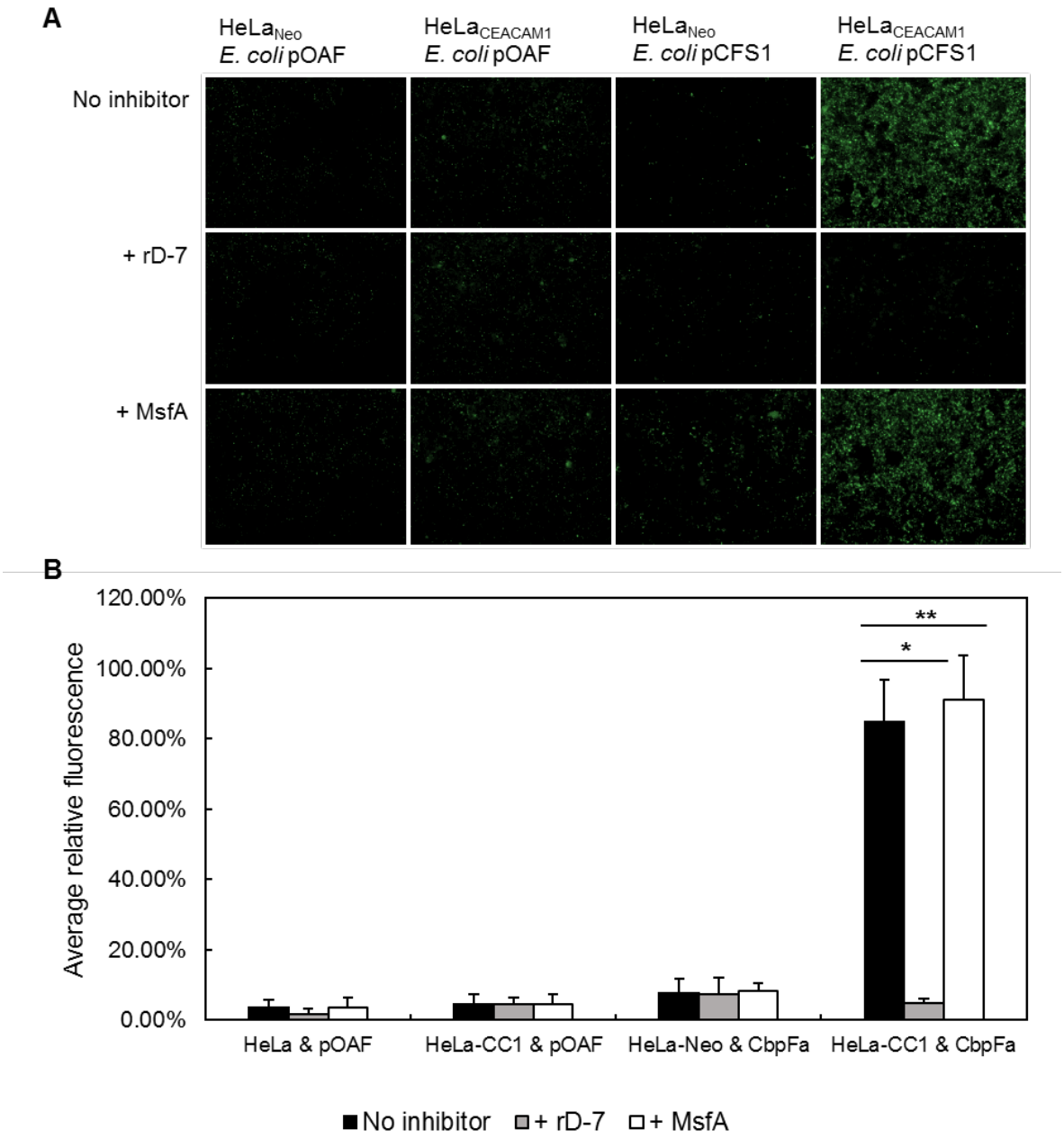


Figure 4.10 | **Inhibition of cellular adhesion between surface-expressed CbpFa and CEACAM1.** A cell adhesion assay with the addition of inhibitor peptides with MsfA and rD-7, which were incubated with the eukaryotic cells for 30 min at 37 °C prior to addition of bacteria. **A)** Sample image set from adhesion assay. **B)** Average relative fluorescence from wells. One-way ANOVA followed by a post-hoc Tukey HSD test yielded significant ($p < 0.0001$) differences for comparisons with any negative control to both uninhibited conditions between CbpFa and CEACAM1 (******); and with the inhibited rD-7 condition compared to the uninhibited ones (*****). There was no significant difference between either positive condition. Key: *E. coli* pOAF – *E. coli* BL21 (DE3) pLysS pOAF (control bacteria); *E. coli* pCFS1 – *E. coli* BL21 (DE3) pLysS pCFS1 (CbpFa; **TABLE 2.3**).

4.3 CbpF CEACAM1 Binding Epitope Elucidation

4.3.1 Alternate TAAs in Adhesion Experiments

To determine possible regions of CbpFa and CbpFb that are binding to CEACAM1, the binding profiles of other similar TAAs were determined. To do this the gene loci FN0471, FN0735 and FNP1391 from *Fn* ATCC 25586 and *Fp* ATCC 10953 were cloned into the plasmid vector pOAF using primers as described in **TABLE 2.4**. The successfully cloned plasmids were subsequently transformed into *E. coli* BL21 (DE3) pLysS cells. Adhesion assays were then carried out using CEACAM1-expressing HeLa cells and the TAA-expressing bacteria. As the peptide fragment rD-7 was shown to inhibit CEACAM1 binding by CbpFa, this was included in the assays to detect any CEACAM1-binding inhibition.

FIGURE 4.11 shows the result of the adhesion assay. CbpFa (FN1499) was included as a positive control and behaved as expected, binding to CEACAM1-expressing cells. Both FN0735 and FNP1391 showed little binding to the HeLa cells and no observable CEACAM1-specific binding. Interestingly, FN0471 bound very well to HeLa cells; however, CEACAM1-specific interactions could not be seen due to the high background. This suggests FN0471 is binding to a different receptor on HeLa cells. At this point, there is no evidence for what this receptor could be. To confirm that there were no CEACAM1-specific interactions with FN0471, western blots were performed on raw bacterial lysates expressing whole protein (not shown). These blots were overlaid with CEACAM1-Fc as used before. The blots came up positive for CbpFa (FN1499) and negative for FN0471, FN0735 and FNP1391, cementing the evidence for it not being able to bind to CEACAM1.

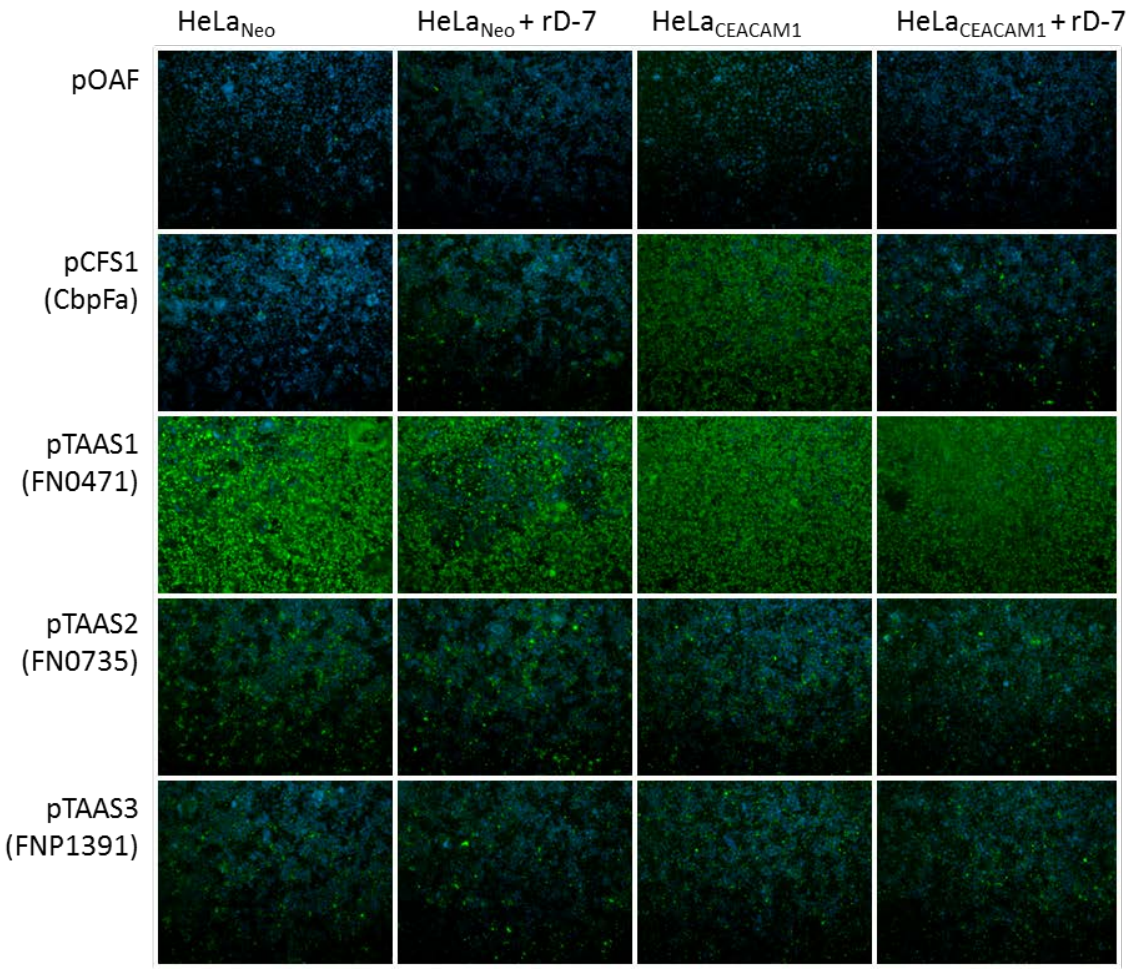


Figure 4.11 | **Alternate trimeric autotransporter adhesins adhesion to CEACAM1.**

The above figure shows an adhesion assay conducted using *E. coli* cells expressing different TAAs on their surface and if any of these can facilitate adhesion to CEACAM1-expressing HeLa cells. Five bacterial strains were used harbouring plasmids as described in **TABLE 2.3**. The peptide rD-7 was included to help distinguish specific and non-specific interactions as well as HeLa cells not expressing CEACAM1 (HeLa$_{Neo}$). Two TAAs from *Fn* ATCC 25586 (FN0471 and FN0735) and one from *F. polymorphum* ATCC 10953 (FNP1391) were tested. In all three cases there were no specific CEACAM1 interactions observed, however, the FN0471-expressing strain bound indiscriminately to all HeLa cells used, indicating that this protein does bind to a human surface receptor. Additionally, FN0735- and FNP1391-expressing bacteria also bound to all HeLa cells, though not to as great an extent as FN0471. This image set is an example from one of three biological repeats.

4.3.2 Rational Design of CbpF mutants

Assuming CbpFa and CbpFb utilise a similar binding epitope on CEACAM1, the region responsible for adhesion on CbpF could be theorised using additive and subtractive alignment methods. The two sequences, in addition to other direct CbpFa homologues, were aligned and the matching regions scored. These hits were subsequently aligned against other TAAs from *Fn* that do not bind CEACAMs (FN0471 and FN0735) in a subtractive manner to remove background sequence hits. The scores for each of the regions of CbpFa are shown in **FIGURE 4.12**. This yields several possible regions in CbpFa that could be responsible for binding to CEACAM1. The residues from CbpFa (FN1499) 148-179, 254-274 and 305-332 are the top three regions that emerge post completing the subtractive alignments. Using this information, deletion mutants and fragments containing these regions were created to experimentally determine the more important sections.

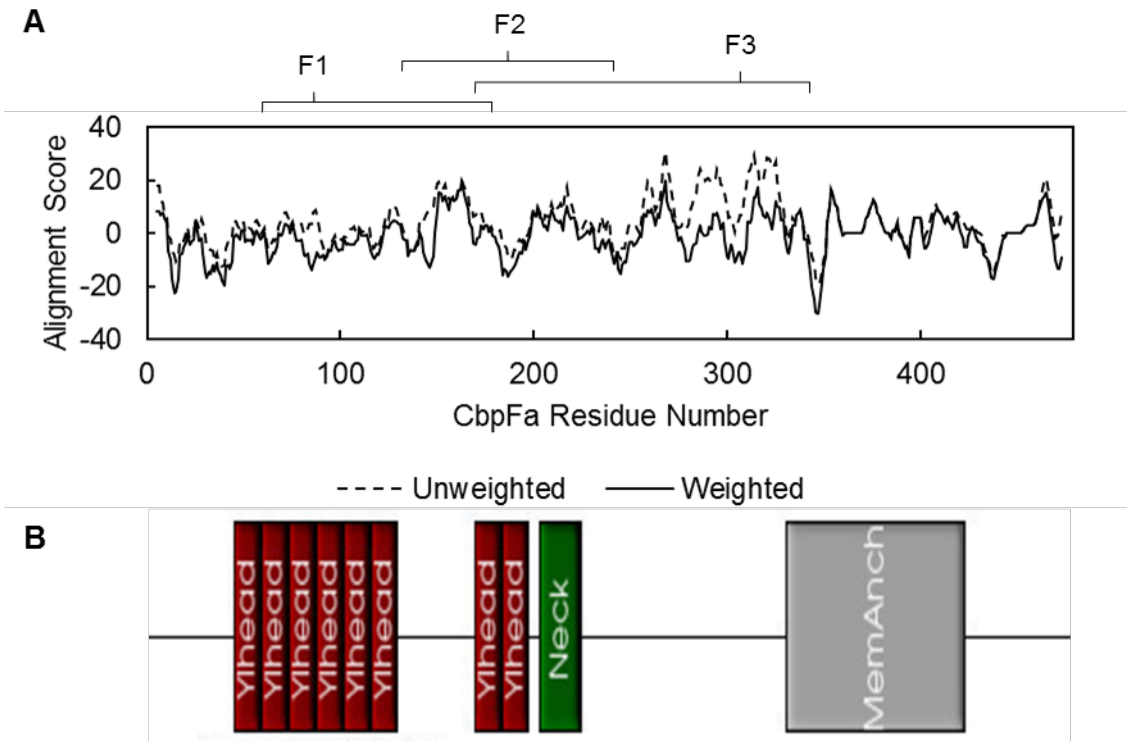


Figure 4.12 | **CbpF sequence propensity.**

A) Sequential 10-residue regions of CbpFa (from *Fn* ATCC 25586) were aligned against and scored against two CbpFb proteins (from *F. oralis* sp. nov. 2B3 and *F. animalis* R5001) and CbpFc (from *F. ovarium* sp. nov.). These regions were also aligned against three other TAAs: FN0471, FN0735 (both from *Fn* ATCC 25586), and FNP1391 (from *Fpoly* ATCC 10953). The two alignment groups were averaged and the non-CbpF group score subtracted from the CbpF score. The net and weighted scores were plotted, where regions greater than 0 are more similar to other CbpFs than any other *Fusobacterium* TAA. The Needleman-Wunsch alignment algorithm (194) was used and alignments were scored using BLOSUM62 (195). Weighted scores produce a lower score for increased variation in the positive alignments. Indicated are the positions of three fragments (F1, F2 and F3) described in **SECTION 4.3.3**. **B)** The domains of CbpFa were annotated using the daTAA (Domain Annotation of Trimeric Autotransporter Adhesins) software package (97).

4.3.3 Fragment production

Initially, CbpFa was split into four distinct domains, each containing different regions spanning the entire extracellular domain. Primers were designed to generate the protein regions 22-128 (pCFR2), 128-180 (pCFR3), 180-235 (pCFR4) and 214-330 (pCFR5) from the FN1499 gene (see **TABLE 2.4** for primers used), and following PCR, were cloned into

pOPINE (**TABLE 2.3**). The resulting vectors were sequenced and checked for any mutations. Full details of the cloning procedure used can be found in the Material and Methods section. All four constructs were made successfully and BL21 (DE3) pLysS cells were transformed with the plasmids for expression.

Multiple conditions were trialled initially using a range of temperatures and incubation periods on a small scale (< 100 ml cultures). A sample of cell suspension was taken post expression and SDS-PAGE loading buffer (**TABLE S 1**) was added and heated for 10 min at 95 °C. The sample was then run down a 4-20 % SDS-PAGE gel and the was either stained with Coomassie or the proteins transferred to a nitrocellulose membrane before performing a Western blot using the polyclonal anti-CbpF sera (PA5154 day 72 or day 56).

Unfortunately, very little to no expression was seen for all four constructs. When scaled up to 1 l cultures with protein being purified using the small-scale native purification method, no protein could be observed when trying to detect CbpF polyclonal antibody binding. Equally for the CEACAM1-F_C conjugate, no binding could be detected. This is indicative of misfolding, at least for the domain involved in binding. From this, the constructs were redesigned and engineered for the pMAL expression system (NEB) that fuses Maltose-binding protein (MBP) to the N-terminal domain. They were also increased in size with a greater overlap between the three resulting constructs: CbpFa residues 40-190 (F1; pCFM2), 120-240 (F2; pCFM3) and 180-331 (F2; pCFM4). The position of these constructs displayed in **FIGURE 4.12**. Immunodot blots were also used this time to increase sensitivity as the proteins would not be denatured and detected in their native form. Protein could be easily detected with the anti-MBP antibody that was supplied with the pMAL expression kit (NEB). **FIGURE 4.13** shows expression of protein with the anti-MBP antibody detecting all three fragments. On this occasion, binding was observed between the proteins and anti-CbpF polyclonal serum, though very weak in the case the CbpFa 120-240 fragment. Yet still, no binding was seen when blotting against CEACAM1-F_C, suggesting the protein fragments may not be completely folded correctly.

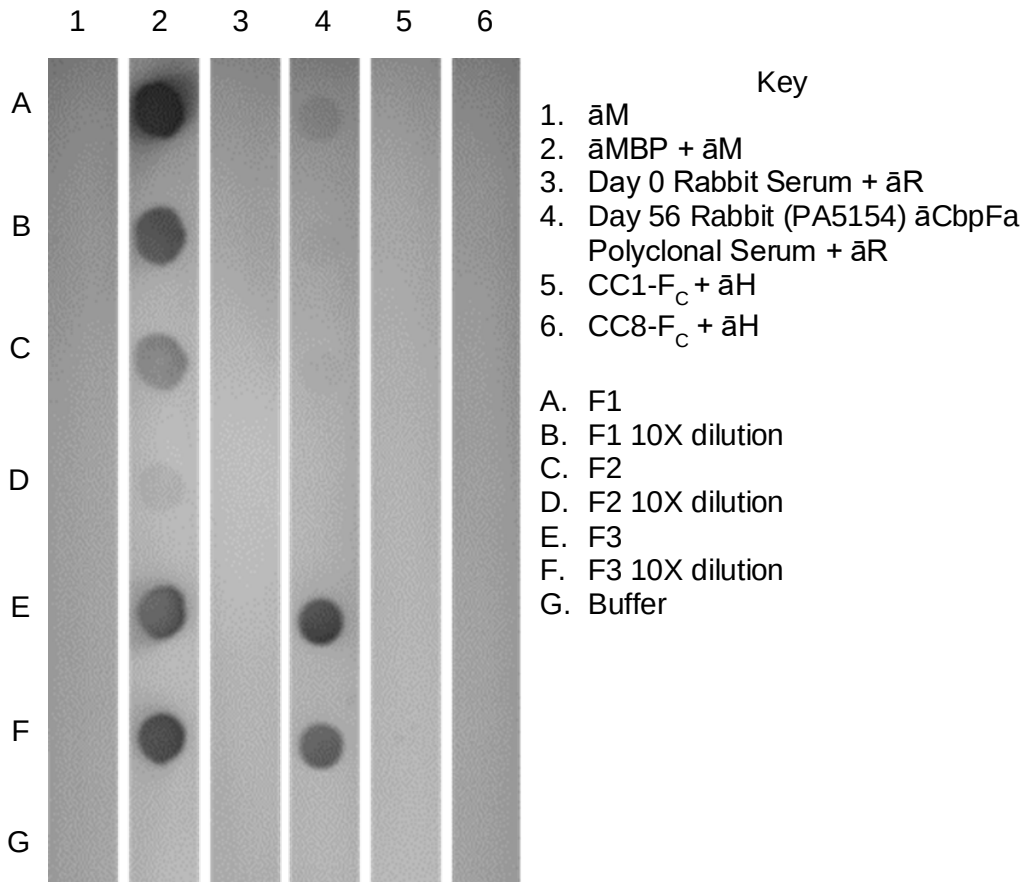


Figure 4.13 | **Dot blot of MBP-tagged CbpFa fragments.**

Rows A-F represent differing protein fragments with a corresponding 10X dilution beneath, detailed in the key. Each column uses a different antibody with a relevant negative control adjacent to it. Proteins were adjusted to 1 and 0.1 μg·ml^{-1} in PBS before loading 100 μl to each position. A buffer only control was also carried out in row G. All proteins were detected successfully using āMBP, though it is evident that F2 is not at the same concentration as the other two fragments, but it was still detectable. āMBP – anti-MBP primary antibody; āM – anti-mouse IgG-AP; āR – anti-rabbit IgG-AP; āH – anti-human IgG-AP; F1 – fragment 1 CbpFa 40-190; F2 – fragment 2 CbpFa 120-240; F3 – fragment 3 CbpFa 180-331.

Because the detection of CbpFa 120-240 fragment was very weak with the polyclonal antibody, it could be due to it lacking the immunodominant epitopes that could bias other domains more greatly. It could also be due to the biological unit not correctly assembling

from steric hindrance from the MBP linked on the N-terminus – though there is a 10-residue long asparagine linker which should prevent this occurring.

4.3.4 Deletion mutants

In addition to creating fragments of protein, some truncation mutants were designed that lacked particular regions of the protein. The two regions designed were CbpFa Δ148-179 and CbpFa 22-283. To create CbpFa Δ148-179, two PCRs were conducted to generate the exterior regions (from residues 22 and up to 330) with XhoI sites added to the deletion ends and pOPINE addons on the exterior DNA ends. The two PCRs were then cut with XhoI and ligated using T4 DNA ligase before inserting into pOPINE using In-Fusion® to create pCFR6 (**TABLE 2.3**). Primers used to create vectors pCFR6 (CbpFa Δ148-179) and pCRF7 (CbpFa 22-283) are described in **TABLE 2.4**.

Unfortunately, the CbpFa 22-283 failed to express, however the Δ148-179 mutant did show some expression and so some initial preliminary tests were carried out on purified protein. The first of which uses the polyclonal antibody against CbpFa to indicate the immunodominant epitopes are exposed, though further tests will need to be carried out to confirm protein folding. The protein was then directly compared to pure extracellular CbpFa in a Western blot using the recombinant fusion protein CEACAM1-4C1-hIgG1 F_C as the primary detector and anti-human IgG-AP conjugated antibody as the secondary detection agent from which the blot was developed (see Methods).

FIGURE 4.14 shows the polyclonal antibody against CbpFa (PA5154) successfully bound to CbpFa Δ148-179, whereas CEACAM1-F_C failed to demonstrate any observable binding, whereas native protein bound both as expected. The results of this experiment are preliminary as more tests are required on purified CbpFa Δ148-179 to confirm the correct folding of all the regions.

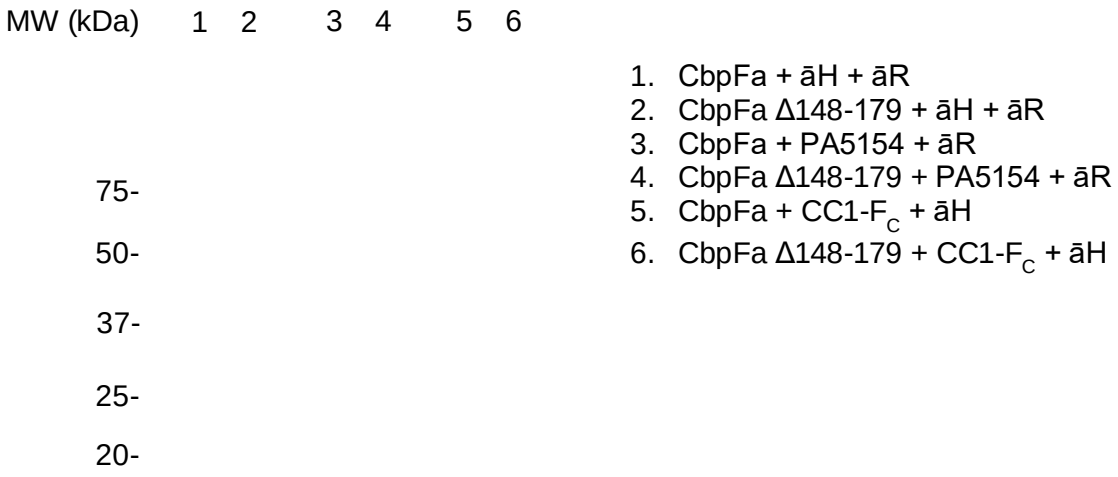


Figure 4.14 | **CbpFa Δ148-179 mutant preliminary CEACAM1 binding results.**

Western blots of purified CbpFa 22-330 and CbpFa Δ148-179 against anti-CbpFa polyclonal antibody (PA5154) and CEACAM1-4C-hIgG1 F_C (CC1-F_C). Lane 4 shows that the polyclonal antibody can bind the deletion mutant and it does run smaller than native protein which is to be expected. There is no observable binding to CEACAM1 with the Δ148-179 mutant though (lane 6) where native protein is behaving as expected (lane 5). A negative control was carried out using anti-human and anti-rabbit secondary antibodies only. Additional bands at higher molecular weights in lanes 3 and 4 can also be seen, which correspond to dimeric and trimeric forms of the protein that have not been fully dissociated. āH – anti-human IgG; āR – anti-rabbit IgG.

4.3.5 Truncation Mutants

Further to the use of fragment and gapped mutants of CbpFa, we created some mutants that would be suitable for surface expression in pOAF. These were sequential truncations originating from the head domain through to the start of the β-barrel. These mutants have the advantage that they will be constrained at the C-terminus by the β-barrel, which should help with correct formation of the biological unit, which may not always be possible in solution. The mutants designed encoded the following CbpFa residues: 111-479, 180-479, 214-479, 293-479 and 329-479 (see **TABLE 2.4** for the primers used).

The successfully cloned regions were transformed into *E. coli* BL21 (DE3) pLysS cells. The cells were grown to mid-log phase and were induced with 1 mM IPTG and grown O/N at 37 °C as previously explained. Adhesion assays were then carried out as before with CbpFa

and the bacteria stained and assessed for CEACAM1-specific binding. The results of these adhesion assays failed to show any specific interactions between any of the truncated protein constructs and CEACAM1-expressing HeLa cells. However these results were preliminary and the experiment had not been optimised, so further tests should still be performed.

4.4 CEACAM1 Mutant Screen

4.4.1 Premade Mutants

To investigate which residues on CEACAM1 are important for binding, several point mutations were made on the N-terminus of CEACAM1. A previous study designed and produced a set of mutants which are as follows: Y34A, Y34F, Y34S, G47L, I91A, I91L, I91T, Q89N and V96A (133). These mutants were designed rationally and were shown to interfere with UspA1 and Opa interactions with CEACAM1 (133, 135). The mutants and the native protein used in this case were from a shorter construct of CEACAM1 containing the N, A1 and B domains, lacking the A2 domain, that CEACAM1-4C1 contains, and were fused to hIgG1-F_C. The CEACAM NA1B product is a natural splice variant known as CEACAM1-3. This shorter variant should not interfere with binding as the A2 domain is not needed for pathogen receptor interactions, but the native CEACAM1-3-F_C form was used as a positive control in these experiments.

Each mutation confers a change on the predicted receptor-binding surface of CEACAM1 and **FIGURE 4.15** shows the alignment of the CEA-family N-terminal domains highlighting where the mutations occur. As previously mentioned, each mutant influenced UspA1 binding with some mutants binding more strongly to UspA1 and some completely knocking out binding and these will be compared later with the effects the equivalent mutants had on CbpFs.

4.4.2 Novel CEACAM1 Mutants

Since CbpFa and b both bind to CEACAM1 and CEA, common residues that overlapped on the binding domain were identified which were unique to CEACAM1 and CEA and not in other non-binders, such as CEACAM3, 6 and 8. Two key residues were identified using a sequence alignment of the N-terminal IgV domains of all CEA family members (**FIGURE 4.15**). The residues identified were a phenylalanine at position 29 (F29) and a glutamine at position 44 (Q44). The combination of these two amino acids is only found in CEACAM1 and CEA, although F29 is also found in CEACAM3 whereas Q44 is not found in any other CEACAM.

To test the importance of these residues, mutants were constructed that altered one of these single amino acids to one found in other CEACAM variants. These mutants and their alternate CEACAM variant-like are as follows: F29I (CEACAM6), F29Y (CEACAM7), F29R (CEACAM8), F29G (mouse CEACAM1b), Q44L (CEACAM3 and CEACAM6), Q44R (CEACAM6 and CEACAM8), and Q44E (mouse CEACAM1b). The mutants were engineered into a CEACAM1-3-hIgG F_C construct, similarly to the previous mutants. The detailed construction of these mutants is described in the Methods section.

4.4.3 CEACAM1 Mutant Binding Results

An ELISA was performed using purified CbpFa and b and all the CEACAM1 mutants stated. Protein was added at 1 pmol per well and incubated O/N at 4 °C in carbonate buffer before blocking with BSA (see Methods). 15 fmol of each CEACAM1-3-F_C point mutant, native CEACAM1-3-F_C, CEACAM1-4C1-F_C and CEACAM1-A1BA2-F_C (CC1 ΔN-F_C) were added to each well and incubated for 3 hrs. A BSA-only well was also kept for each condition as a secondary negative control. All wells were then overlaid with an anti-human IgG-AP secondary antibody at a total 10 µg per well and incubated for 1 hr. All incubations were carried out at RT and washed between each step (as described in the Materials and Methods). The ELISA was developed for 5 hours at 37 °C and the absorbance was measured at 405 nm. The results for which are detailed in **FIGURE 4.16**.

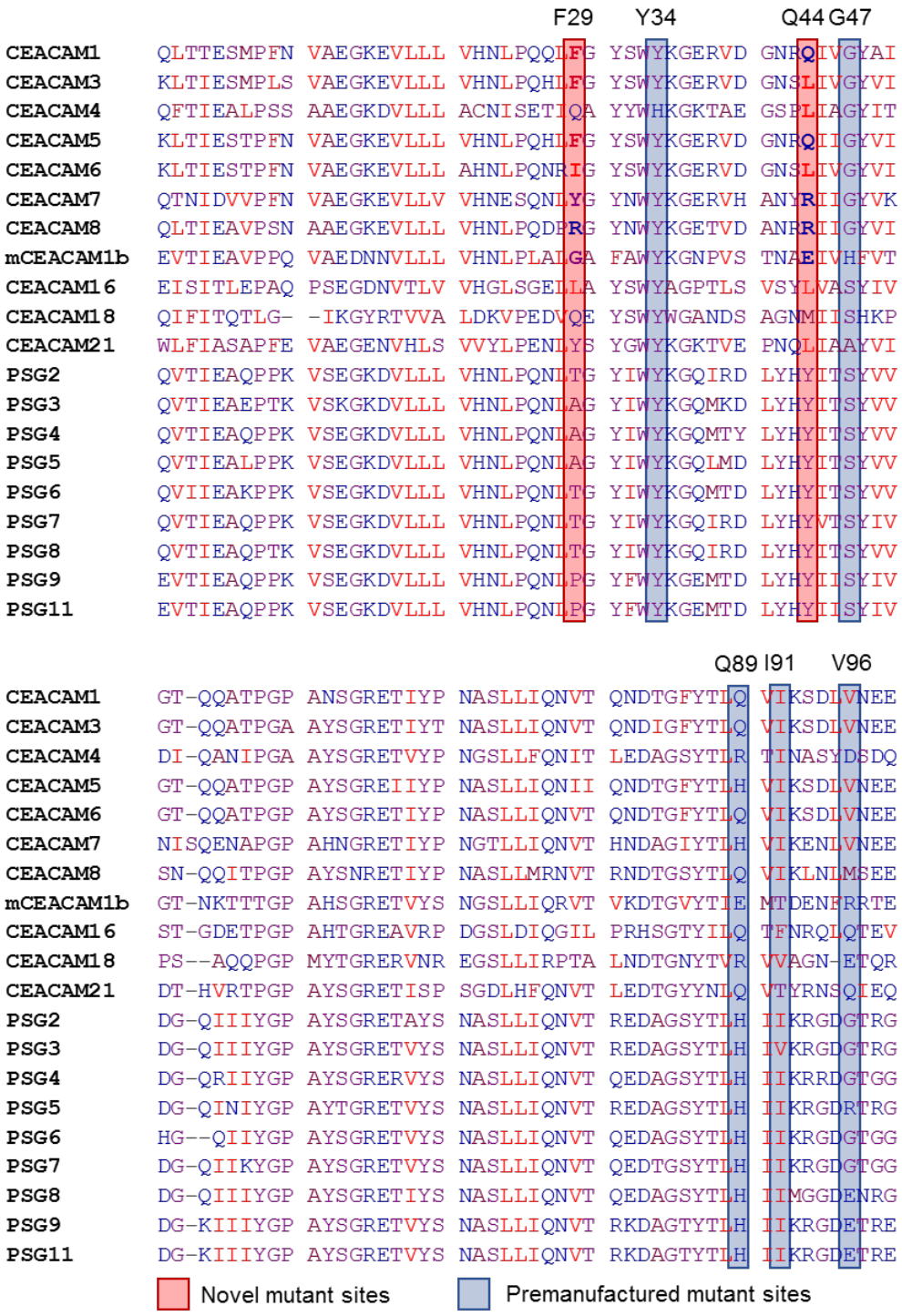


Figure 4.15 | **Sequence alignment for the IgV-like domains of all human CEA-family members.** The sequences of all human CEA-family members IgV-like domains were aligned using Clustal Omega (172). The partial IgV-like domain of CEACAM20 and the C-terminal IgV-like domain of CEACAM16 were excluded. The N-terminal of murine CEACAM1b (mCEACAM1b) was also included for comparison. Sites at which mutations were made in the recombinant CEACAM1-F_C conjugate are also highlighted. See **APPENDIX B** for CEACAM numbering conventions.

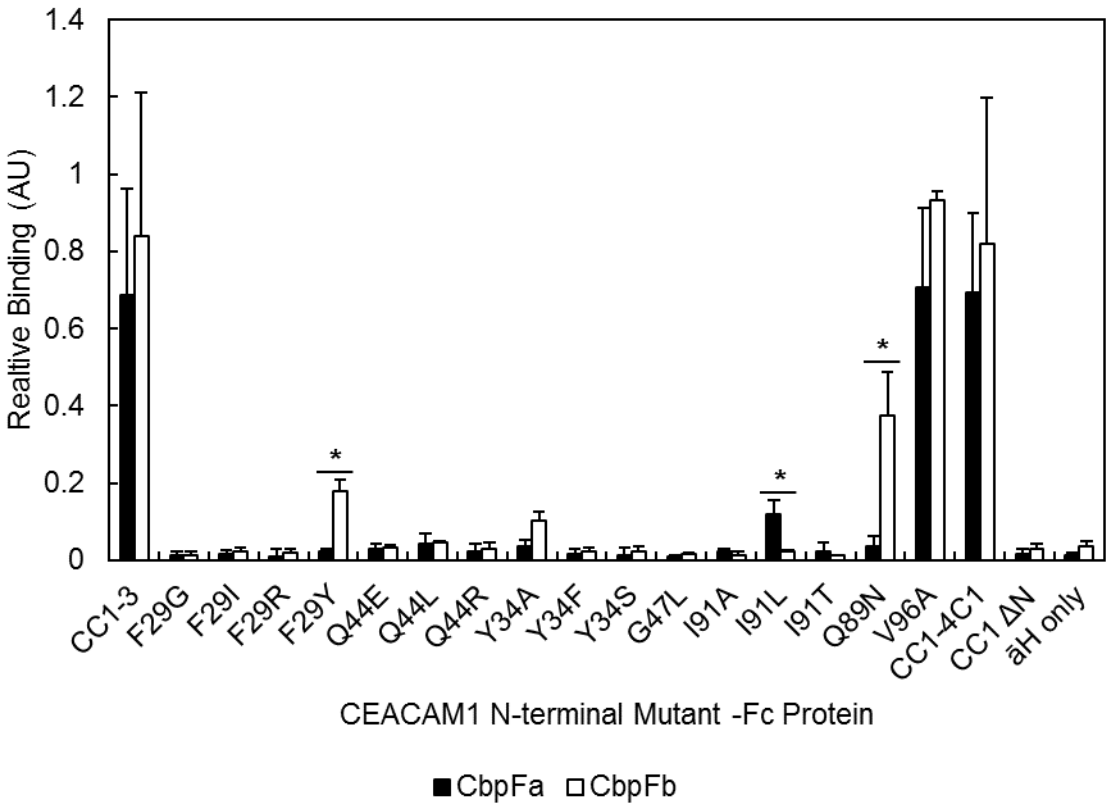


Figure 4.16 | **CbpF-CEACAM1 mutant binding assay.**

This graph shows the results of an ELISA examining the effects that different CEACAM1 N-terminal mutations have on adhesion to CbpFa and CbpFb. Equimolar amounts of each protein (3 pmol) were overlaid overnight followed by blocking with BSA and incubation for 3 hours with 15 fmol of mutant CEACAM1-3-hIgG1 F_C (CC1-3-F_C). In addition, CEACAM1-4C-F_C (CC1-4C) and CEACAM1-A1BA2-F_C (ΔN) were used as positive and negative controls respectively as well as a BSA only control. Anti-human IgG-AP (āH) was used to detect binding. The ELISA was developed for 5 hrs at 37 °C using the SigmaFast® kit. A two-way ANOVA followed by a post-hoc TukeyHSD test was used identify significant differences conditions (N-terminal mutant: $F(19) = 42.86,\ p < 0.0001$; CbpF protein $F(1) = 7.87, p < 0.01$). Individual T-tests were carried out for the marked paired conditions and were found to be significantly different (* $p < 0.05$). $N = 3$.

Examining **FIGURE 4.16** and post statistical analyses with ANOVA, no significant difference between both positive control conditions, CEACAM1-3 and CEACAM1-4C1, was observed, as expected. The V96A mutant also had a negligible effect on CbpF adherence and was not significantly different.

Changing Y34 in all cases had a negative impact on protein binding, where Y34A showed little though not significant, binding to CbpFb but Y34F and Y34S knocked detection down to basal levels for both CbpFs. Similarly, altering I91 had a detrimental effect on protein interactions, where I91L showed some adhesion to CbpFa only.

Interestingly, Q89N bound reasonably strongly to CbpFb when compared to the other mutants, however, showed no observable interactions with CbpFa indicating a specificity not present in CbpFb.

The F29G, I and R mutants all had a substantial impact on the binding potential for both proteins reducing it to basal levels. Interestingly, the F29Y mutant still shows some interaction with CbpFb but reduces it completely for CbpFa. Other than this, both CbpFa and b do not appear to bind any of the other mutants with Q44 appearing critical for CEACAM1 or CEA adhesion as this residue is unique to both these CEACAMs throughout the CEA family of proteins. In a separate preliminary experiment (**FIGURE 4.17**), rD-7 was shown to bind all Q44 mutants, in fact potentially having a higher affinity toward Q44E than native, though Q44L and Q44R both reduce it.

The results from this study are largely preliminary with a mutant concentration limitation not allowing to study saturating conditions, where only very small quantities of mutant were added per well.

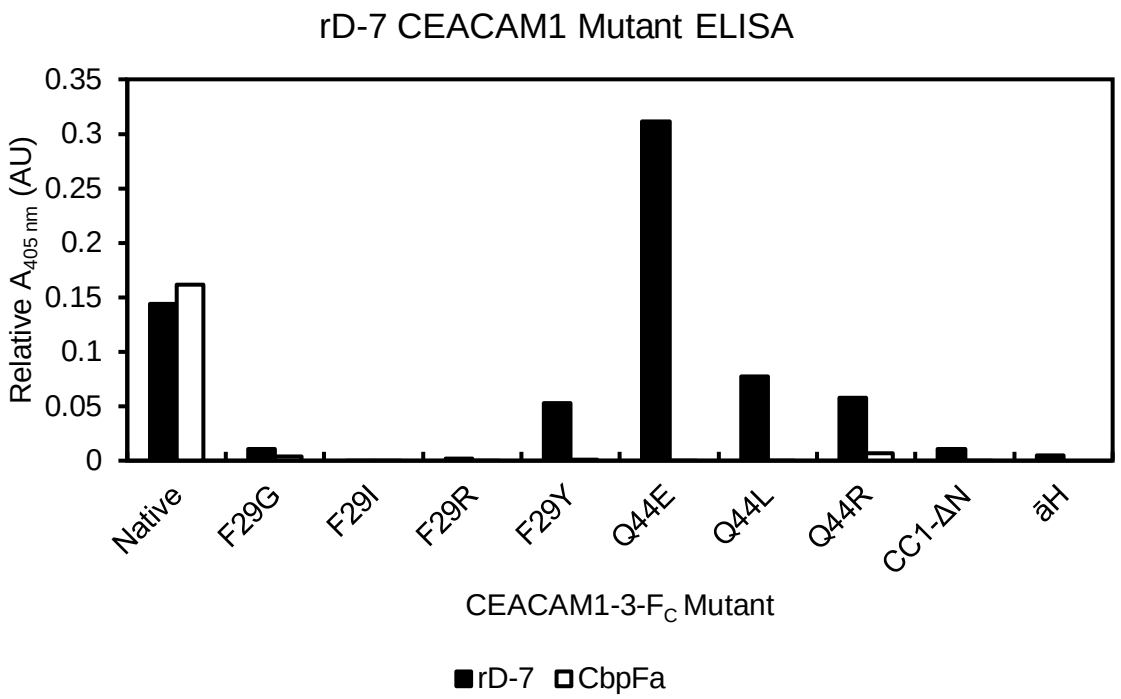


Figure 4.17 | **rD-7 interactions with CEACAM1-3 IgV mutants.**

The above figure is a preliminary study (n=1) examining rD-7 interactions with the novel CEACAM1-3 N-terminal IgV mutants. The experiment was carried out with identical parameters to the one detailed in **FIGURE 4.16** with CbpFa included as a control. As can be seen, rD-7 was able to bind all Q44 mutants to some degree and even bound the Q44E mutant with a higher affinity than with native protein. Changing F29 however knocked out binding, though the F29Y mutant still showed some adhesion. CC1-ΔN – CEACAM1-A1BA2-F_C. āH – anti-human IgG.

4.4.4 CEACAM1 N-terminal Mutant Molecular Dynamics

To assess the validity of the CEACAM1 N-terminal domain mutants, spatiotemporal molecular dynamics (MD) was employed. In addition to validating structural stability, we wanted to compare any feature differences displayed by the mutants compared to the native form in an attempt to explain observed differences in CbpF-binding. For example, the mutants could show less flexibility than the native, with this potentially reducing the availability of CbpF-binding site.

MD was performed using the GROMACS (version 5.0) package (154, 155). The crystal structure of the IgV-like domain of CEACAM1 (PDB ID: 4WHD) was used as a template from which the relevant side chains were mutated and separate models were created. The system setup, energy minimisation, equilibration and production MD steps are described in detail in the Methods section. In brief, a truncated octahedral unit with periodic boundary conditions (PBC) containing the protein was solvated with water and NaCl (0.1 M) where the salt ions were adjusted to yield neutral conditions. Following energy minimisation, temperature and pressure equilibration, 20 ns of production MD was run (See **APPENDIX F** for longer runs of wild type protein.). The PBC were removed and the system centred on the protein backbone. Following this, the RMSD (root-mean-square deviation) of the protein backbone was calculated for the 20 ns of MD. Both the original crystal and energy minimised structures were used to calculate RMSD, where the original structure was used as a reference to confirm there were no major structural artefacts introduced following minimisation. The graph of these values is shown in **FIGURE 4.18-A**. Five rounds of MD were performed for each structure to confirm the validity of the results, where all cognate simulations displayed highly similar features.

A large deviation can be seen early in the MD for the native CEACAM1 where there is a considerable jump at 6 ns. This jump is not observed in most of the other simulations, though the Q44R mutant does make this jump and sooner, just after 3 ns. **FIGURE 4.18-B** shows this deviation where three structures are shown from three points throughout the MD

simulation (0, 10 and 20 ns). One loop on the binding face (around the residue T52 between the C' and C'' sheets) folds outward, were it remains for the remainder of the simulation. Conversely, the F29Y mutant, for example, does not display this outward folding and neither do the other mutants (**FIGURE 4.18-C**).

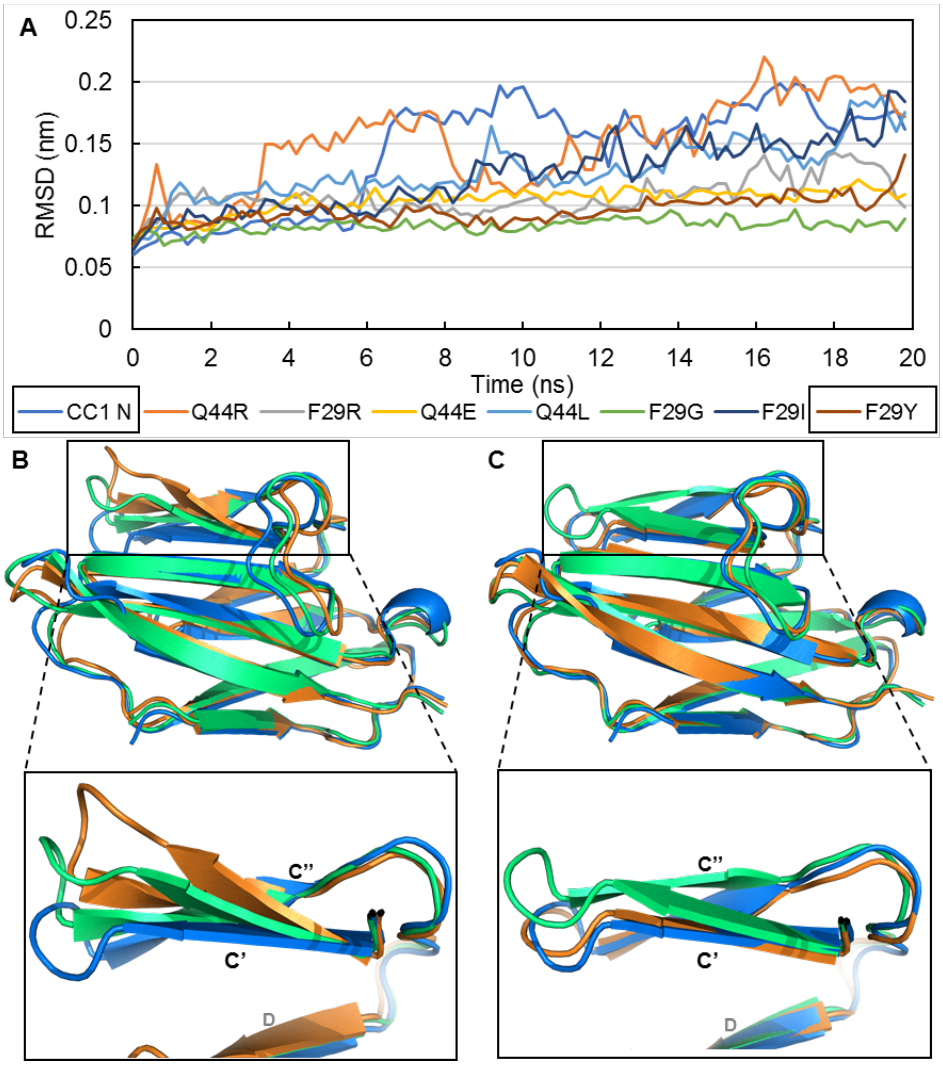


Figure 4.18 | **Molecular dynamics of CEACAM1 mutants.**

A) The backbone RMSD (against the post-minimised state) values are from 20 ns of molecular dynamics simulation. The base protein was the N-terminal crystal structure of CEACAM1 (4WHD; CC1 N), which was mutated to produce the mutant models listed above. For clarity, the graph was smoothed by taking averages at 20 ps intervals. A continued RMSD below 0.3 nm is generally considered stable, and every protein examined here remained well below that value. **B)** Three frames from native CEACAM1 and **C)** from the F29Y mutant simulations. Blue – 0 ns; orange – 10 ns; green – 20 ns. All frames were superposed against a common reference.

4.5 Discussion

4.5.1 CEACAM binding is Species Dependant

Unlike other TAAs from *Fusobacterium*, CEACAM-binding capable TAAs have only been identified in four species: *nucleatum*, *vincentii*, *animalis* and *oralis* sp. nov. Cellular binding has hitherto only been observed in these species, though an *oralis*-like CEACAM1-binding homologue has been identified in a previously unclassified *animalis* strain, *Fa* strain, CAG:649; it is unknown whether this strain can bind CEACAM1. However, three of the clinical strains (R5001, R15792 and R30927) fall into the same clade, but only two (R5001 and R15792) exhibit the same homologue and were shown to bind CEACAM1, with R15792 not consistently displaying adhesion. No other homologues have been identified elsewhere in *Fp, Fw*, *Fh* or *Fperio*. The various clinical strains that bound CEACAM all derived from *Fn*, *Fv*, *F oralis* sp. nov., *F. ovarium* sp. nov. or one of the two *Fa* strains.

The CEACAM-binding TAA candidates identified are all found upstream of the *nik*-operon, which is responsible for nickel transport across the membrane. Downstream of this in *Fn* ATCC 25586 is a transposase gene, suggesting this region was at one point a mobile genetic element, though divergence of the TAAs into two very distinct classes suggests it is no longer mobile. Moreover, the presence of a homologue in only a few *Fa* strains suggests an ancestor of this clade obtained the gene and this group diverged separately from the main clade in *F. animalis*, though some of the strains in this subgroup have lost this gene as it may not have imposed an evolutionary disadvantage for these species. Conversely, it is found ubiquitously within *Fn*, *Fv* and *For* species, which suggests it is required for these particular species.

Interestingly the newly classified *F. ovarium* sp. nov. species could also bind CEACAM1, though using a much smaller protein. As the putative *cbpFc2* gene product is far removed from the main two clusters of CbpFs (**FIGURE 4.6**), in addition to not existing adjacent to the *nik*-operon, it will be disregarded as the CEACAM-binding candidate. Further evidence for this can be seen in **FIGURE 4.3**, where the size of the CEACAM1-binidng protein is

substantially smaller than the other proteins, whereas the CbpFc2 protein would be expected to be the same size as CbpFa. The CEACAM1-binding protein band more appropriately matches the CbpFc1 protein (approximately 100 amino acid residues shorter than CbpFa). Therefore, CbpFc will henceforth be used to refer specifically to the CbpFc1 protein from *Fov*.

With regards to the potential impact on disease for *F. animalis*, R5001 was isolated from the placenta; R30927 was isolated from a maternal blood culture following a preterm birth; MJR7757B was isolated from the vagina; CAG:649 isolated from GI tract – human gut metagenome (female with Crohn's Disease); and R15792 isolated from knee pus (male with infected knee). From this limited data, it is hard to draw meaningful conclusions about this subgroup of *Fa*, though it is worth noting that these species may be more linked to preterm birth and pregnancy complications from two of the cases listed; however, the presence of the CbpF protein does not correlate with these disease phenotypes as the R30927 has no discernible *cbpF* gene. So, as previously stated it is possible the CbpFb protein is not required for these *Fa* strains, though further evidence will be required to assess the importance of CbpF in *Fa*.

4.5.2 CbpFs Bind CEACAM1 and CEA Through Interactions with Specific Residues

Experiments examining CbpFa and b binding, both on cells and solution, show that the only receptors, for which they can bind, is CEACAM1 and CEA (CEACAM5), with no definitive evidence seen for CEACAM3, 6 or 8 or mouse CEACAM1b. This, however, does not mean they cannot bind to other CEACAMs or other CEA-related receptors such as PSGs. Nonetheless, binding CEACAM1 specifically has the advantage of being able to potentiate local immunosuppression via the ITIM domain contained within the intracellular region of CEACAM1 (127, 196), while avoiding the immune-activating CEACAM3 receptor found on granulocytes, which could lead to bacterial phagocytosis (134).

Other studies have examined ITIM signalling through the TIGIT receptor and how this may prevent natural killer (NK) cells from targeting tumour cells that *Fusobacterium* is bound to (45). This particular study focusses on the role of the Fap2 protein, but does not look into CEACAM1 effects, although mentioning it momentarily. If a similar effect could be induced by CEACAM1-CbpF interactions, this would provide more evidence for an active role of *Fusobacterium* in CRC progression for instance.

As shown in the ELISA for comparing different point mutants on the predicted N-terminal binding interface, certain residues are important for maintaining interactions to CbpF. For example, Q44 appears to be a key residue involved in these interactions. This residue is shared only between CEACAM1 and CEA not appearing in any other CEA family member (**FIGURE 4.15**). As mutating this residue destroys binding to either CbpFa or CbpFb, it can be hypothesised that there is a direct interaction between CbpF and Q44 that is responsible for specificity.

The F29 residue is another critical residue for CbpF interactions, again only appearing in CEACAM1, CEA, and CEACAM3. As neither CbpF can adhere to CEACAM3 (as determined by protein-only assay), this is not the only residue that is conferring adhesion. Moreover, mutating this amino acid to a tyrosine still allows CbpFb to bind, though slight, but not CbpFa. Interestingly, this F29Y change is found naturally in CEACAM7, which, unfortunately, was unable to be tested in any of the experiments carried out so it would be an informative experiment to perform in the future. However, as CEACAM7 lacks the Q44 residue, where it has an arginine in its place, it will likely be unable to bind either CbpF as all Q44 mutants knocked out adhesion

Additionally, mutating Y34 had a large impact on the ability for CbpFs to bind. This is likely a larger structural change that would alter this domain making it more flexible. Y34 is highly conserved among all the CEA-family members also appearing in the IgV-like domains of PSGs (**FIGURE 4.15**); this suggests it is a vital residue to confer normal function and structure of these proteins.

The mutation with the least impact on CbpF adhesion was the V96A mutant where the difference was not significant between either native CEACAM1 variant. This strongly implies that this residue is not involved in adhesion and does not form part of the binding interface. The only other mutant that is predicted to be on the binding surface of CEACAM1 that did not completely reduce adhesion, at least for CbpFb, was Q89N. As previously explained, there is a difference in affinity between the interactions of CbpFa and CbpFb with CEA where CbpFb binds more strongly to it than CbpFa, whereas, both bind to CEACAM1 equally. One of the main differences between CEACAM1 and CEA is the amino acid residue Q89 where CEA has a histidine at this position. As changing this normally polar amino acid to a slightly positively charged amino acid (histidine in CEA) or normally negative residue (glutamic acid in CEACAM1 mutant) does not impact CbpFb as much as CbpFa, this residue can be theorised to not play as much of a role in binding to CbpFb. Conversely, it seems more important for maintaining CbpFa affinity and when changed to glutamic acid, completely removes the ability to adhere. **FIGURE 4.19** shows the difference between the N-terminal domains of CEACAM1, 5 and 6. Highlighted are the locations of some of the key amino acid mutations that affect CbpF-CEACAM1 interactions.

In the I91L mutant, binding was retained slightly for CbpFa, however when changing this residue to alanine or threonine, the affinity was reduced completely. This mutant was the only example of CbpFa being able to bind more strongly than CbpFb, where there was no observed adhesion to any of the I91 mutants. Like Y34, I91 is highly conserved in the CEA family, appearing in all the native CEACAMs examined here, but it is an interesting feature that only CbpFa can bind, albeit slight, to a mutant of this residue.

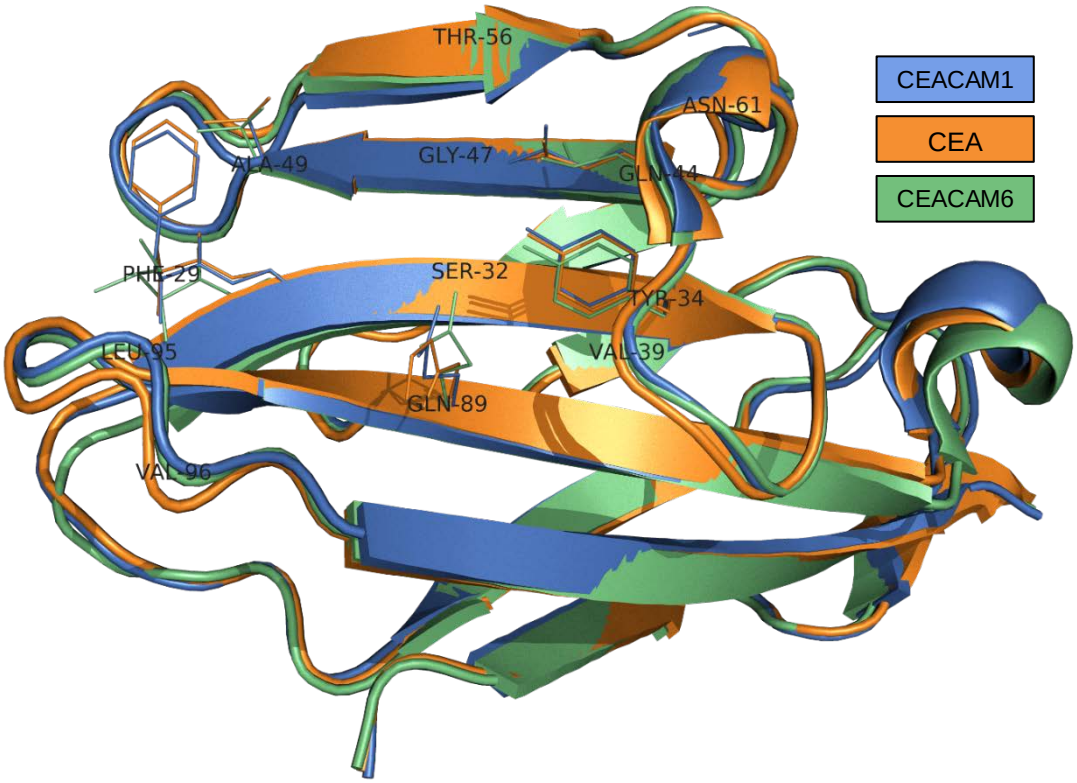


Figure 4.19 | **Structural superposition of the N-terminal domains of CEACAMs.**

The structures of CEACAM1, CEA (CEACAM5) (139) and CEACAM6 (137) N-terminal domains (PDB IDs: 4WHD, 2QSQ and 4Y8A respectively) have been structurally aligned using superposed. Residues on the predicted binding interface have been labelled according to the residue on CEACAM1. Residues of interest have been highlighted with their sidechains represented as lines. The residues of the binding interface are highly similar between CEACAM1 and CEA, with two notable differences at residues 89 and 49 changing from glutamine to histidine and alanine to valine from CEACAM1 to CEA respectively. CEACAM1 and CEA also share certain residues not present in the CEACAM6 N-terminal domain such as F29 and Q44 which are replaced by isoleucine and leucine respectively. This residue is not present in the structures of other CEACAMs where in CEACAM6 it is replaced with an isoleucine.

Given the time and resources, this study would have examined other CEACAM1 mutants as well as mutants from CEA and CEACAM3. For example, it would be informative to see if mutating the CEACAM3 residue L44 to glutamine could allow binding of CbpFs. Q44 is the only unique sidechain to CEACAM1 and CEA, as CEACAM3 also contains the F29 residue, unlike all other members of the CEA family. Likewise, it would be interesting to see if mutating the H89 on CEA to glutamine would increase binding affinity for CbpFa, as this is the primary difference on the CFG face between CEACAM1 and CEA. Notably, Q89H is also a naturally occurring single nucleotide polymorphism (SNP) for CEACAM1 found within the human genome. Nonetheless, the information gathered in this study provides an insight to the possible molecular interactions between CEACAM1 and CbpFs.

Chapter 5: Structural Analysis and Modelling of CbpFs

5.1 Introduction

To help build an understanding of the molecular interactions between CbpFs and CEACAMs, structural information was needed to strengthen hypotheses generated from functional data. Structures of TAAs such as UspA1, BpaA, YadA and SadA have significantly aided the modelling of important molecular interactions with their respective receptor-targets (102, 103, 107, 108, 135). The HopQ virulence factor from *H. pylori* was even co-crystallised with the IgV-like domain of CEACAM1 providing a detailed model for protein-protein interactions and what epitopes would be important in vaccine design (140).

Prior to this study, there was no structural information pertaining to CbpFs. As these proteins belong to the trimeric autotransporter adhesin family of proteins, some assumptions can be made regarding the overall topology. The most straightforward of which is the β-barrel, which is highly conserved among all TAAs and is extremely similar in CbpFs. In addition to this, there should be at least two other regions in the extracellular domain, consisting of a coiled-coil stalk extruding from the membrane anchor followed by a series of constraining head domains. However, other substructures could exist, for example it could have multiple CC regions or altering head-neck topologies, with examples detailed in **TABLE 1.1**, such as the case for SadA which has a complex domain organisation over the complete fibril (108). However, it could be a much simpler situation, such as YadA, only having one of each of the core components of stalk, neck and head domains (103).

As these proteins originate from a species not closely related to other species with known TAA structures, there will likely be some structural nuances specific to *Fusobacterium*, especially considering the restricted codon bias imposed by the inherently low GC % content.

5.2 Computational prediction

5.2.1 Sequence Analysis

In addition to attempting to experimentally solve the structures of CbpFa and b, computational predictions of the secondary and tertiary structures were made using a variety of software. By performing these structure and domain predictions, we hoped this could aid in solving crystal structures through molecular replacement or provide the overall model, should experimental determination have failed. In addition to providing structural information, the analyses could help quickly identify similarities and differences between the different classes of CbpFs between species, such as domain possession or deletion to identify the key domains for function.

Initially the overall topology of the proteins was determined using daTAA (Domain Annotation of Trimeric Autotransporter Adhesins) (97), which finds regions of homology with a dictionary of known TAA domain sequences. **FIGURE 5.1** shows the annotation of the proteins according to daTAA. From examining the annotations provided by daTAA, there appear to be gapped regions where unknown and potential novel folds exist, with a common region in both CbpFa and CbpFb between the two YadA-like head clusters. The second smaller YadA-like cluster also has a different motif from that of the N-terminal cluster. The N-terminal YadA-like head cluster consists of conserved 14-residue repeat sequences and the motif has a well-defined fingerprint as shown in **FIGURE 5.2**. Each one of these repeats constitutes one turn of an O-shaped β-roll, as described in detail in **SECTION 1.2.2**. The second cluster of YadA-like domains that were annotated cannot be resolved to a common consensus sequence, though the glycine at the twelfth position, also in the first cluster, remains conserved in all four sequences.

The closest neighbour to the N-terminal group of head domains, with a solved structure, is UspA1 from *M. catarrhalis*, however it contains one extra residue per turn of the O-shaped β-roll. The neck region that follows is most like that of Hia from *H. influenzae* (75 % identity) and BpaA from *B. pseudomallei* (73 % identity). By aligning the tetradecamer repeat

sequence from CbpF to UspA1, we can predict the internal facing residues of the domain. The highly conserved SSAFG from positions 8-12 likely form the inner core of the trimeric biomolecule with the bulky phenylalanine residue forming hydrophobic interactions within the central core.

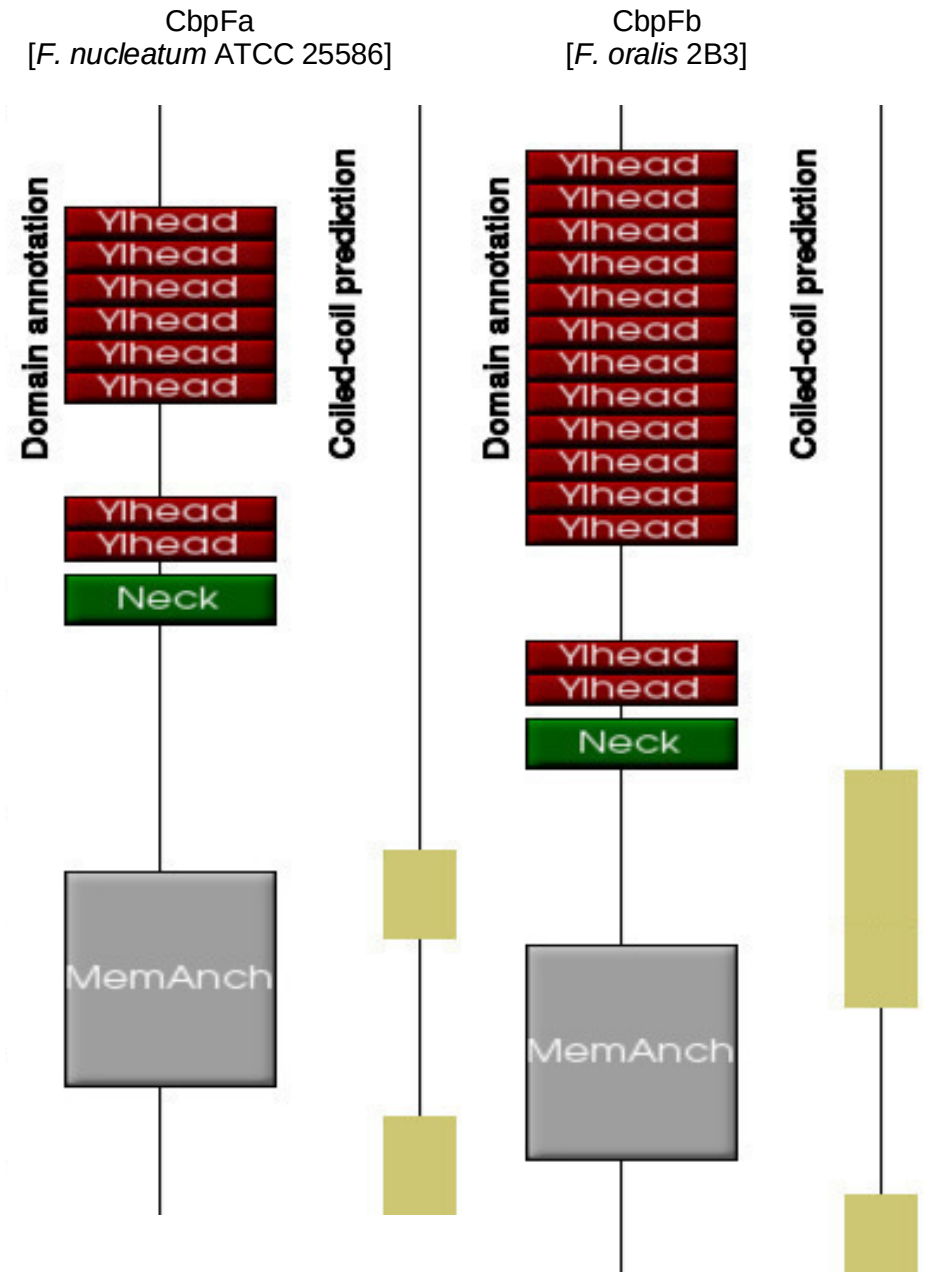


Figure 5.1 | **Domain Annotation of CbpFa and b using daTAA.**

The software package daTAA (Domain Annotation of Trimeric Autotransporter Adhesins) was used to highlight conserved regions on both CbpFa and CbpFb using a domain dictionary containing all the known folds found within TAAs. Three known domains were detected, which were YadA-like head groups (YIhead), a neck domain and the membrane anchor. The domain-dictionary alignment lookup also finds potential coiled-coil domains which are shown adjacent to the domain annotations. Additionally, it appears daTAA missed an extra YadA-like group off the end of the first cluster, which was identified by looking at the sequence manually.

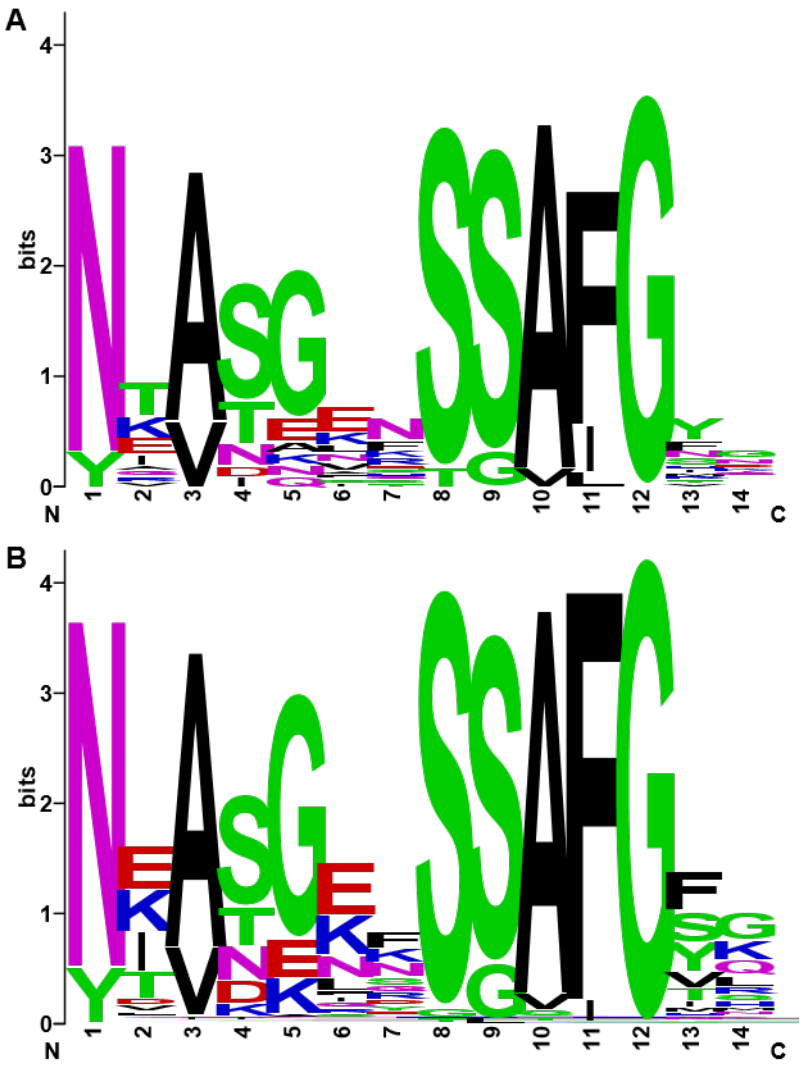


Figure 5.2 | **CbpFa and b YadA-like head repeat sequence logo.**

This shows the propensities for each amino acid residue of the tetradecamer repeat sequences within the initial cluster of YadA-like head domains in **A)** CbpFa (7 domains) and b (12 domains) and **B)** all identified CbpF proteins within *Fusobacterium* spp (314 domains combined from 44 proteins). The most conserved residue is at position 12, which was glycine and is present in all repeats from every species, including CbpFc. Positions 1, 3, 8, 9, 10 and 11 were all highly conserved and replaced by a similar amino acid if different. The exception to this is in the final group within the cluster where there is a tyrosine at position 1 in all CbpF proteins. Positions 4 and 5 are also weakly conserved, whereas residues 2, 6, 7, 13 and 14 show no clear preference. Sequence logos were generated using WebLogo (197). A detailed breakdown of the domains is shown in **FIGURE S 6**.

The region of the protein that has the largest (for CbpFa) unknown motif is the domain directly downstream from the neck that should comprise the stalk section of the protein. For CbpFa, this region does not have any indication that it would form a coiled-coil, when using the programs MARCOIL and PCOILS. Likewise, CbpFb is not predicted to have a coiled-coil region here either, however, daTAA does predict a CC region between the neck and the membrane anchor (**FIGURE 5.1**). **FIGURE 5.3** shows the CC prediction for four CbpF proteins including CbpFa, b and c. The stalk region should start at residue 236 for CbpFa directly after then neck, however, no CC inclination can be observed until much further downstream directly before the start of the membrane anchor. Moreover, the region that has some CC prediction is largely not included in the constructs used in this study as it is likely buried within the β-barrel of the membrane anchor and it is known not to be necessary for CEACAM1 interactions. The main reason for MARCOIL and PCOILS not predicting a CC, is likely due to the presence of a helix-breaking proline residue in the centre of this region, although a caveat to this is the fourth protein examined in **FIGURE 5.3**, where CbpFc does not have this poly-glycine-proline motif.

FIGURE 5.4 shows the sequence alignment from the end of the first YadA-like head cluster to the start of the membrane anchor. The beginning of the stalk region starts at residue 236 on CbpFa and 306 on CbpFb. There is one insightful gap that CbpFb has when compared to CbpFa where it is missing 23 amino acids and then there is also a short 2-residue gap which then follows on CbpFa. As the average residues per turn on an CC alpha helix is approximately 3.5 with 7 possible registers total, it would be expected that the helix-breaking proline would follow the same register in both proteins, therefore only gaps of a multiple of 7 would be allowed. When combining the two gapped regions on the alignment, the net difference in length is 21, equivalent to 3 full register cycles, or approximately 6 helix turns. This then poses another question regarding what follows, which is currently unknown where running a protein BLAST against the Protein Data Bank (PDB) database yielded no homologous regions. When examining the full alignment of all identified putative CbpF

proteins (**APPENDIX D**), there is one exception to this CC register re-complementation. The presumed CbpFa from *Fn* R28400, a strain known to bind CEACAM1 (**FIGURE 4.2**), only has a net gap length of 20, where there is an extra glycine present at residue 262. All other CbpFa and CbpFb proteins have a net gap of 21. CbpFc has a large deletion in this whole domain, lacking the glycine-rich region followed by a proline, so is more likely to be all CC.

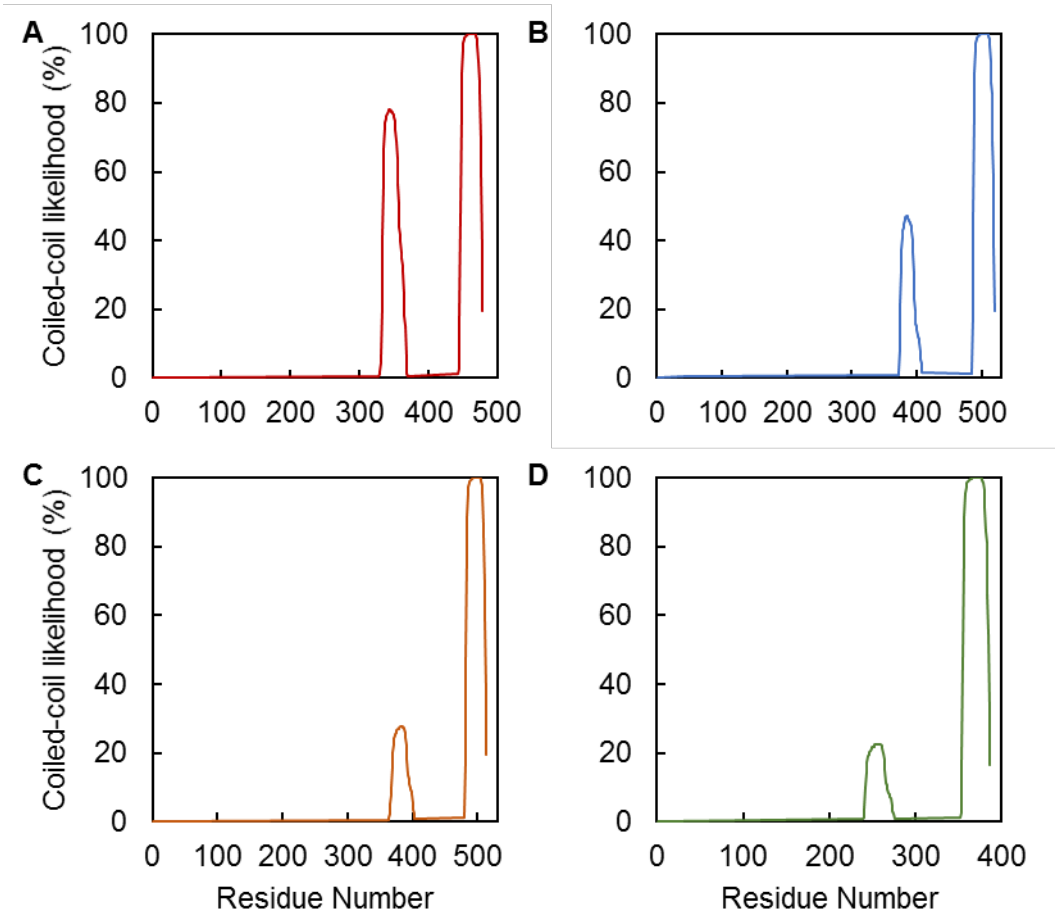


Figure 5.3 | **CbpF MARCOIL coiled-coil prediction**

MARCOIL was used to predict the coiled-coil likelihood for each region of four of the CbpF proteins including **A)** CbpFa [*Fn* ATCC 25586], **B)** CbpFb [*For* 2B3], **C)** CbpFb [*Fa* R5001] and **D)** CbpFc [*Fov* R16531]. All four show some propensity for a CC directly prior to the membrane anchor (first peak in all four panels) followed by a strong signal after the membrane anchor region. The stalk region starts after the neck domain at residues 236, 306, 292 and 221 for each protein respectively.

```
CbpFa   GMGEFN-GQYQ YKNEGNNSY MIGNKNKIAS GSDDNFILGN NVHIGGGINN  187
CbpFb   GVGFWNSGSHL YKNEGNNSY MIGNKNKIAS GSDDNFILGN NVEIGAGVQK  257

CbpFa   SVALGNNSTVS ASNTVSVGS STLKRKIVNV GDGAISANSS DAVTGRQLYS  237
CbpFb   SVVLGDGSASG GSNTVSVGS STLQRKIVNV ADGTISATST DAVTGRQLYS  307

CbpFa   GNGIDTAAWQN KLNVTRKND YKDANDIDVN KWKAKLGVGS GGG--GGAPV  285
CbpFb   GDGID------ --------- --------VN KWRTKLGVSS GGGASGGAPG  334

CbpFa   DAYTKSEADNK FANKTDLND YTKKDDYKDA NGIDVDKWKA KLGTG       330
CbpFb   DAYTKSEADNK FTSK----- ----DDYKDA NGIDVDKWKA KLGTG       370
```

Figure 5.4 | **CbpFa and b uncharacterised region sequence alignment.**

This alignment covers the least well-defined region of both CbpFa [*Fn* ATCC 25586] and CbpFb [*For* 2B3] from the end of the first YadA-like head group cluster to the start of the membrane anchor. The stalk region beings at residues 236 and 306 for CbpFa and b respectively.

5.2.2 Model Building

For model building, several different online servers were used, with some using quantum mechanics for creating *ab initio* models and others relying on existing molecular homology. The combination of all these generated models was expected to give an indication whether the structures generated had any validity. The applications used were: SWISS-Model (198), M4T (199), I-TASSER (200), QUARK (201) and RaptorX (202). For QUARK, the input sequence was limited to 200 residues due to the complex nature of its algorithm. All four applications were able to generate models of the head domain, as expected, due to its highly conserved nature.

FIGURE 5.5 shows an example of the models that were produced. It is evident that the models do not represent a complete structure of each protein. Interestingly, the RaptorX program managed to model further down the sequence of CbpFb compared to CbpFa, modelling approximately 100 more amino acid residues. Nonetheless, none of the tested modelling software gave high confidence for any of the modelled regions. In addition, none of the modelling software could produce a structure for the stalk domain. Out of the software used, only QUARK has the ability to build completely *ab initio* models with no prior information (201), however, no reliable structure could be made. As previously mentioned, the stalk

region has no homologues within the Protein Data Bank (PDB) (203), therefore programs that rely on pre-existing templates to produce models will fail to predict this region accurately.

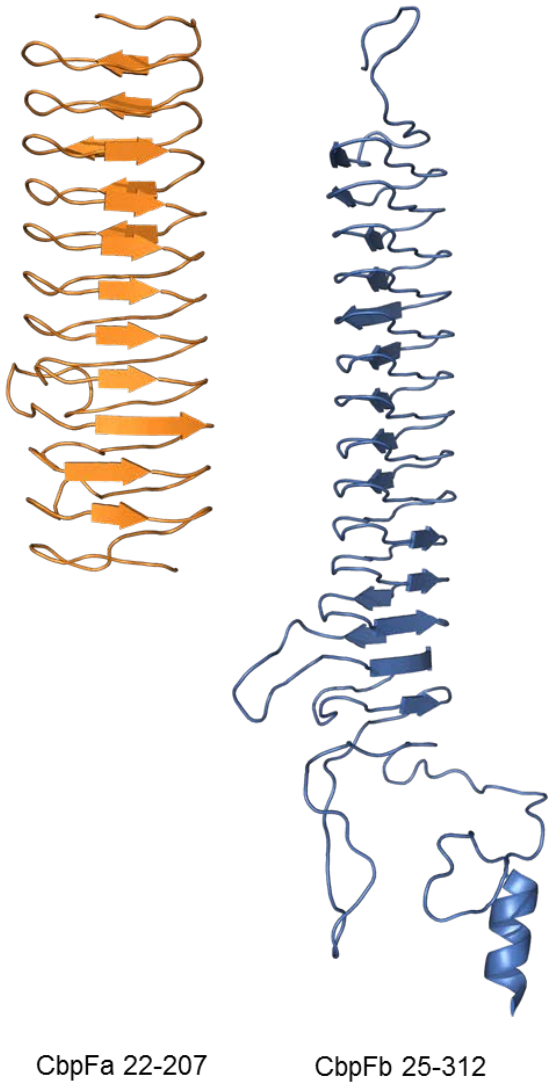


Figure 5.5 | ***De novo* models of CbpFa and CbpFb.**

The models presented were generated by RaptorX and were cropped such that disordered low confidence regions were excluded for clarity. Residues from the signal-sequence to the C-terminus were input constituting approximately 480 amino acid residues. The model lost confidence in both cases following the first sequence of YadA-like head domains, however it managed to model further using CbpFb (up to residue 312; blue) than CbpFa (up to residue 207; orange).

The application CCBuilder (204) was used to artificially build *de novo* coiled-coils from an input sequence. As it is expected there should be a coiled-coil in the stalk domain, in conjunction with other TAAs and because prediction software refuted any existence of one, they had to be guessed. The top scoring models for stability were used to educate the register, as it could exist within 7 different conformations (**FIGURE 1.8**). Other information such as pitch and interface angle were drawn from the CCs belonging to most similar structural homologues to the stalk CbpF such as SadA. This should at least have comparable properties at the N-terminal end of the stalk where the CC joins the neck region; the neck region of CbpF shares high homology to equivalent regions of the SadA and BpaA structures. This was done in an attempt to solve a crystal structure as detailed in the following section.

5.3 Experimental Data

5.3.1 X-ray Crystallography

Recombinant protein was prepared for CbpFa and CbpFb using the pCFR1 and pCBR2 vectors respectively (**TABLE 2.3**). To achieve sufficiently high protein yields for use in protein crystallography, expression conditions were optimised using a variety of *E. coli* expression strains, expression media, induction time, expression time and temperatures. The most optimal expression conditions used are described in detail in **SECTION 2.4.2.** In short, Rosetta2® (DE3) pLacI *E. coli* were grown overnight in LB broth before diluting 1:100 in 8 l autoinduction terrific broth and were subsequently grown at 37 °C for 24 hrs with shaking at 200 RPM. Cells were pelleted by centrifugation and stored at -80 °C prior to protein purification.

The proteins were initially purified using affinity chromatography with Ni-NTA agarose as described in **SECTION 2.4.2**. Elution fractions collected were analysed for protein using SDS-PAGE and Coomassie straining, where fractions containing protein were pooled and concentrated to approximately 5 ml. Any contaminants were then removed using size exclusion chromatography (SEC). See **FIGURE 5.6** for example purification.

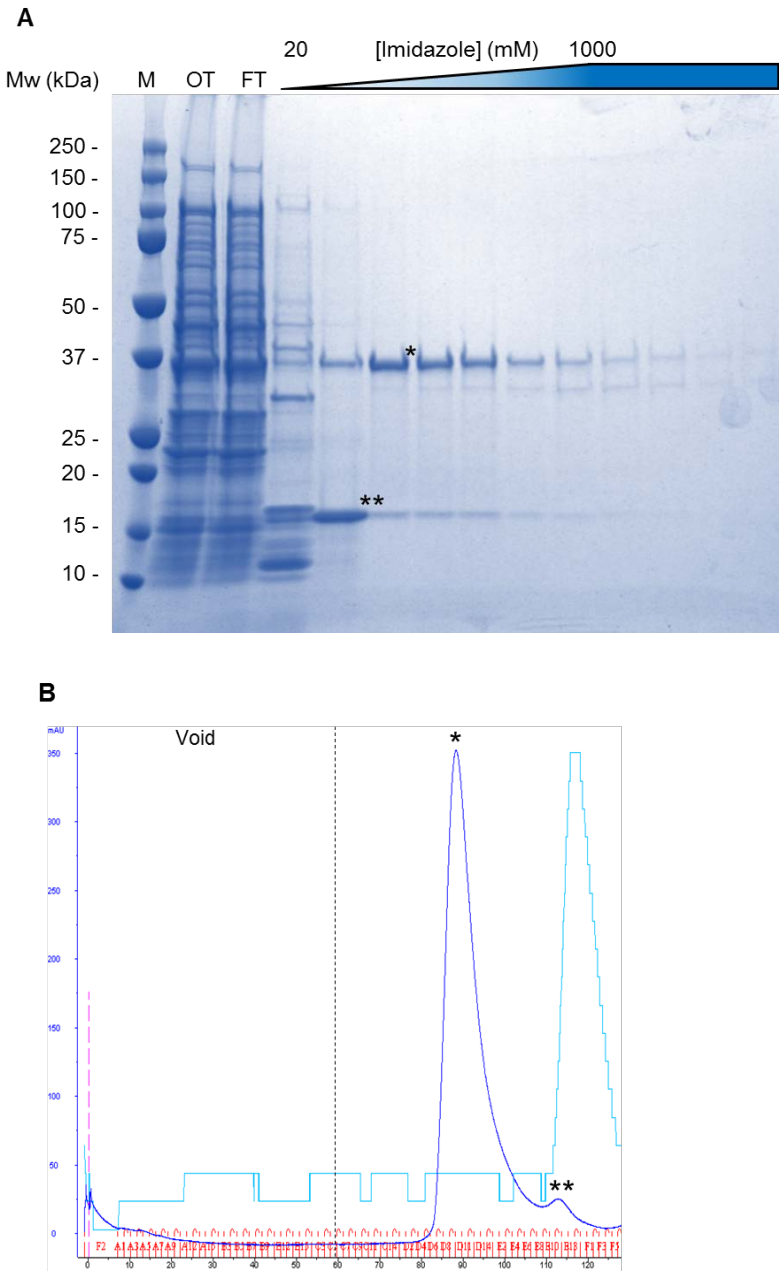


Figure 5.6 | **Example affinity and size-exclusion chromatography for CbpFb.**

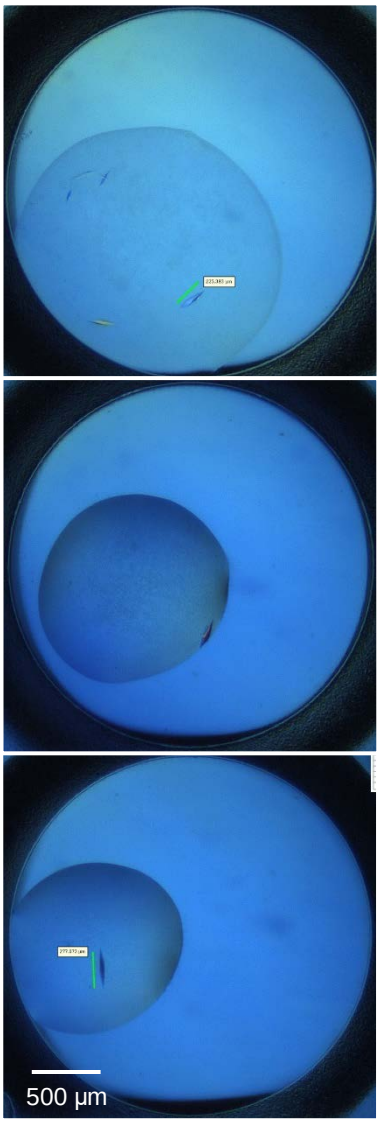
Cell lysate was loaded onto an Ni-NTA resin affinity chromatography column where the column was washed with 20 column volumes of loading buffer, followed by a gradient increase in imidazole concentration to elute bound proteins. Fractions from each stage were loaded onto and SDS-PAGE gel and run at 300 V for 30 min and stained with Coomassie quick stain (**A**). M – Marker; OT – Pre-column sample; FT – flow-through pre elution. Fractions containing CbpFb protein (* expected 36.1 kDa for monomer) were pooled and dialysed and run down a size-exclusion column where it was separated from degraded or contaminating proteins (**; **B** – shows the $UV_{280\ nm}$ absorbance of the fractions over the cumulative volume of buffer passed down the column).

Following SEC and concentration, purified protein was laid on a multitude of 96-well high throughput crystallography screens (sitting drop vapour diffusion; **SECTION 2.9**), with varying drop volume, protein to precipitant ratio and protein concentration. The initial screens used were: JCSG-*plus*™, MIDAS-*plus*™, Morpheus®, Morpheus® II, PACT *premier*™, ProPlex™, HELIX™ and Structure Screen 1+2 (Molecular Dimensions). Any promising crystal hits were then optimised about by varying the well solution component concentrations, pH and protein concentrations.

Unfortunately, CbpFa could not ever be optimised to give crystals suitable for data collection, though further screening and optimisation may be able to yield a crystal worth examining on the beamline. However, CbpFb had a few conditions where it crystallised readily without the requirement to optimise conditions. The most reliable condition contained 0.1 M MES pH 6.5, 0.05 CsCl, 30 % (v/v) Jeffamine® which yielded protein crystals in the two screens that share this condition (JCSG*plus*™ and Structure Screen I+II). The optimal condition had a one-to-one ratio of protein (at 3.2 $mg \cdot ml^{-1}$) to precipitant solution at 200 nl each in the drop and 50 µl precipitant solution in the reservoir. Crystals were visible after incubation at 20 °C for two weeks. Images were captured periodically to monitor growth and samples of positive conditions are shown in **FIGURE 5.7**. Images were captured using visible, cross-polarising and UV light and crystals absorbing strongly in the UV were examined further as these were more likely to be protein crystals.

Crystals were looped and incubated in a buffer-matched cryoprotectant containing up to 30 % (v/v) glycerol (see Methods). Unfortunately, several crystals dissolved or were severely damaged when cryopreserving. However, some did not undergo dissolution – these were crystals from the JCSG*plus*™ F1 (JF1) and MIDAS™ E11 (ME11) wells (see product sheets on Molecular Dimensions for screen details). These crystals were frozen in liquid nitrogen before collecting data on a high-energy X-ray beamline (I04-1 Diamond Light Source). Examination of the crystal at the beamline indicated there were some ice crystals, so the

crystals were washed three times by inserting and removing the loop from the liquid nitrogen stack. Data frames were collected with 360° of rotation and 5 frames per degree.



JCSG*plus*™ F1
200:200 nl protein to precipitant drop
50 µl reservoir volume
Well solution:
0.1 M MES pH 6.5, 0.05 M CsCl, 30 % (v/v) Jeffamine® M-600

ProPlex™ E6
200:200 nl protein to precipitant drop
50 µl reservoir volume
Well solution:
0.1 M sodium cacodylate pH 5.5, 0.1 M calcium acetate, 12 % (w/v) PEG 8000

ProPlex™ G12
200:200 nl protein to precipitant drop
50 µl reservoir volume
Well solution:
0.1 M HEPES pH 7.0, 1 M tri-sodium citrate

Figure 5.7 | **CbpFb optimal crystallisation conditions.**

Shown are images captured with a cross-polarising light filter of conditions that yielded protein crystals of CbpFb. In addition to the conditions shown, the condition from Structure Screen I+II F12, which shares the conditions of JCSG*plus*™ F1 also produced crystals. The pCBR1 plasmid, encoding CbpFb 25-374 [*F. oralis* 2B3] and expressed in *E. coli* Rosetta2 pLacI cells, was used. The stock protein concentration that yielded crystals was 3.2 $mg \cdot ml^{-1}$ (in SEC Buffer A).

The data collected from the JF1 crystal showed strong diffraction with the resolution reaching 2.0 Å from initial analysis. The resulting data were subsequently processed using iMOSFLM (151). The data were integrated, merged and scaled using iMOSFLM and

AIMLESS (205) within the CCP4 software suite (152). A summary of the data and statistics is listed in **TABLE 5.1**. The dimensions of the resulting unit cell, as calculated by POINTLESS (206, 207), were as follows: $a = b = 59.83, c = 497.28, \alpha = \beta = 90.00, \gamma = 120.00$ (where a, b, c correspond to length in Å and α, β, γ correspond to angle in degrees). The space group R 3 2'' (H 3 2; corresponding to a H-centred trigonal unit cell) was determined to be the most probable solution (Laue group probability = 0.998). The wavelength (λ) from the data collection was 0.916 Å. Following cell content analysis, the percent solvent was calculated to be 48.1 % with a Matthews Coefficient of 2.37 and 1 monomer in the asymmetric unit (probability = 1.0000).

Table 5.1 | **Summary of X-ray diffraction data for CbpFb**

[1] $R_{merge} = \frac{\sum_{hkl} \sum_j |I_{hkl,j} - \langle I_{hkl} \rangle|}{\sum_{hkl} \sum_j I_{hkl,j}}$

	Overall (51.36-2.74 Å)	**Highest Resolution (2.87-2.74 Å)**
R_{merge}[1]	0.041	0.317
$N_{observations}$	34717	4633
N_{unique}	9438	1214
$I/\sigma(I)$	12.3	3.4
Completeness	99.3	97.5
Multiplicity	3.7	3.8

Following a BLAST search against the PDB database, several partial homologues to CbpFb were identified with the closest match having a coverage of 84.4 %, with a percent identity of 33.4 %. This protein was a crystal structure of head and neck domain of the UspA1 protein (PDB ID: 3PR7). Details of all the trialled structures are listed in **TABLE 5.2**. Each of these structures was remodelled using Sculptor (208) in the Phenix application suite (153) with varying parameters corresponding to sequence conservation. The models were then sequentially parsed to Phaser (209) for molecular replacement, with one monomer in the unit cell. In addition, cropped and whole models produced by RaptorX, SWISS-Model and M4T were also trialled in molecular replacement. None of the *de novo* coiled-coils built using

CCBuilder, with varying registers, pitch and interface angles, produced a solution with molecular replacement in Phaser.

The model that yielded the highest scoring data post molecular replacement was a sculpted model of the head domain of 3PR7. Electron density was modified prior to statistical chain tracing with BUCCANEER (210) and SHELX (211). The model was then iteratively manually built and refined. Model building was performed in Coot (212) and refinement with REFMAC5 (213). A model including residues 18-175 of the expressed protein could be built before the electron density map could no longer be built into. After multiple phase-build-refine iterations, no accurate structure could be produced, with final Free-R of approximately 0.5.

Table 5.2 | **Molecular replacement template list.**

A BLAST search of the expressed region of CbpFb [*For* 2B3] against the PDB database was performed and the resulting top hits are listed. Each template was subsequently trimmed and edited using Sculptor and each was trialled in molecular replacement.

PDB ID	Length	Coverage (%)	Identity (%)	Similarity (%)
3PR7	302	84.4	33.4	45.0
3NTN	209	58.4	34.0	46.9
3S6L	120	33.5	36.7	55.8
3WP8	169	47.2	33.7	46.2
2YO3	53	14.8	52.8	62.3
3LA9	71	19.8	43.7	52.1
2YO0	72	20.1	45.8	59.7
4USX	35	9.8	57.1	71.4
2XQH	110	30.7	35.5	52.7
3EMO	24	6.7	70.8	83.3
1P9H	48	13.4	39.6	60.4
2YNZ	24	6.7	66.7	75.0
3ZMF	23	6.4	60.9	78.3
2YO2	23	6.4	60.9	78.3

To rule out contaminating proteins or degraded protein, the JF1 crystal was retained, washed and resolubilised in water. The solution was then analysed using liquid-chromatography-mass spectrometry (LC-MS; a service provided by the University of Bristol proteomics facility). The results found no evidence of contaminating proteins and the presence of the majority of the expected protein (69.6 % coverage). The LC-MS results are detailed in **APPENDIX H** (**FIGURE S 7**).

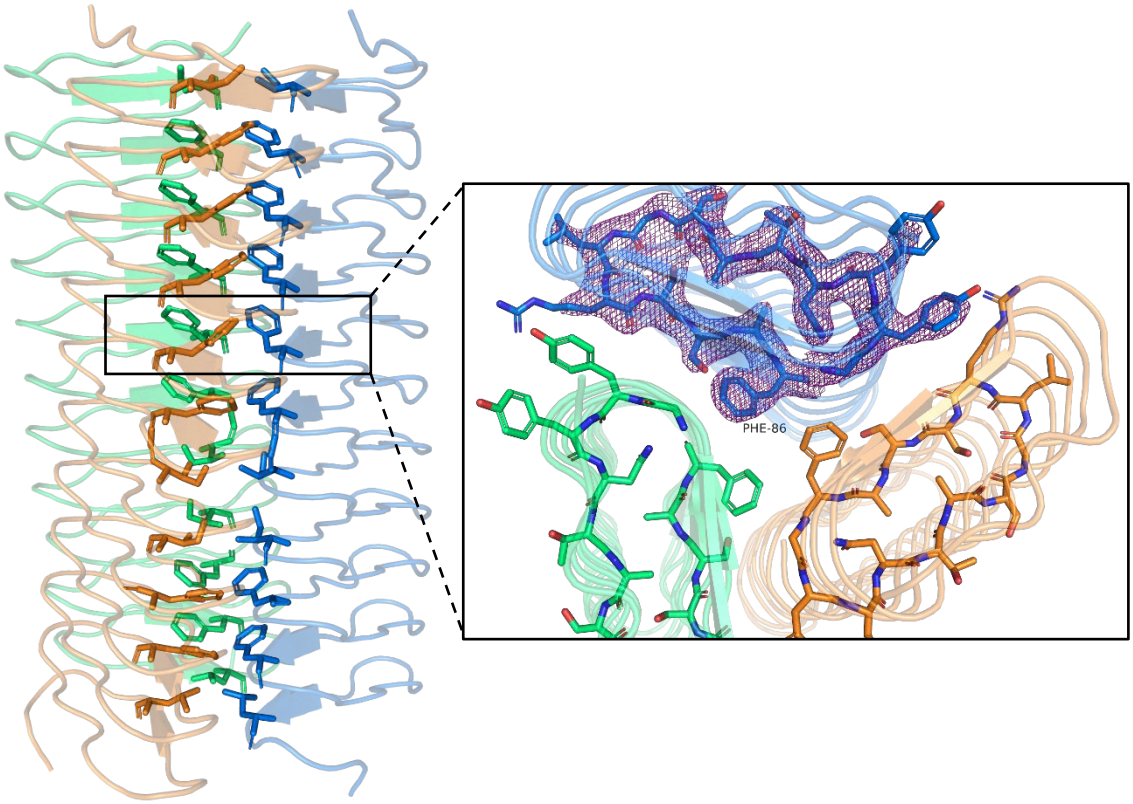


Figure 5.8 | **Best model for CbpFb from X-ray data.**

The above model shows the regions 18-175 of CbpFb that could be modelled post molecular replacement, model building and refinement. Highlighted is the central core consisting mainly of phenylalanine residues, as well as isoleucine and leucine, forming the 'phenylalanine zipper'. One turn of the tetradecameric head domain is shown on the right (residues G73-F86) and the best corresponding map overlaid (2Fo-Fc map with a 1.2 sigma contour).

5.3.2 Circular Dichroism

As CbpFa did not readily crystallise, indirect experimental methods were employed to infer its structural features. Initially, circular dichroism (CD) was used to estimate secondary structure proportions and whether this aligned with computational prediction from the sequence. Three concentrations of CbpFa 22-330, from the *E. coli* Rosetta™ 2 (DE3) pLacI::pCFR1 construct (**TABLE 2.3**), were prepared in SEC Buffer A (**TABLE S 1**) at 1, 2 and 5 µM before collecting CD data with 190 to 250 nm wavelength, with a 0.5 nm interval. Relative ellipticity was recorded and is displayed in **FIGURE 5.9-A**.

Several CD deconvolution and spectra matching programs were used to get an approximation of the relative composition of the protein. A newer algorithm devolved by Micsonai *et al.* (214, 215) was used and the proposed composition is detailed in **FIGURE 5.9-B**. Due to the presence of chloride ions in the preparation buffer, the beta-sheet estimations will not be as accurate compared to the alpha helices. However, the overall estimation aligns well to what would be expected from examining the daTAA result, where the YadA-like head domains contribute toward 40 % of the beta-domains alone. The uncharacterised stalk region would only account for 30 % of the structure, which should be predominantly alpha-helical, which compares to the 23 % estimation.

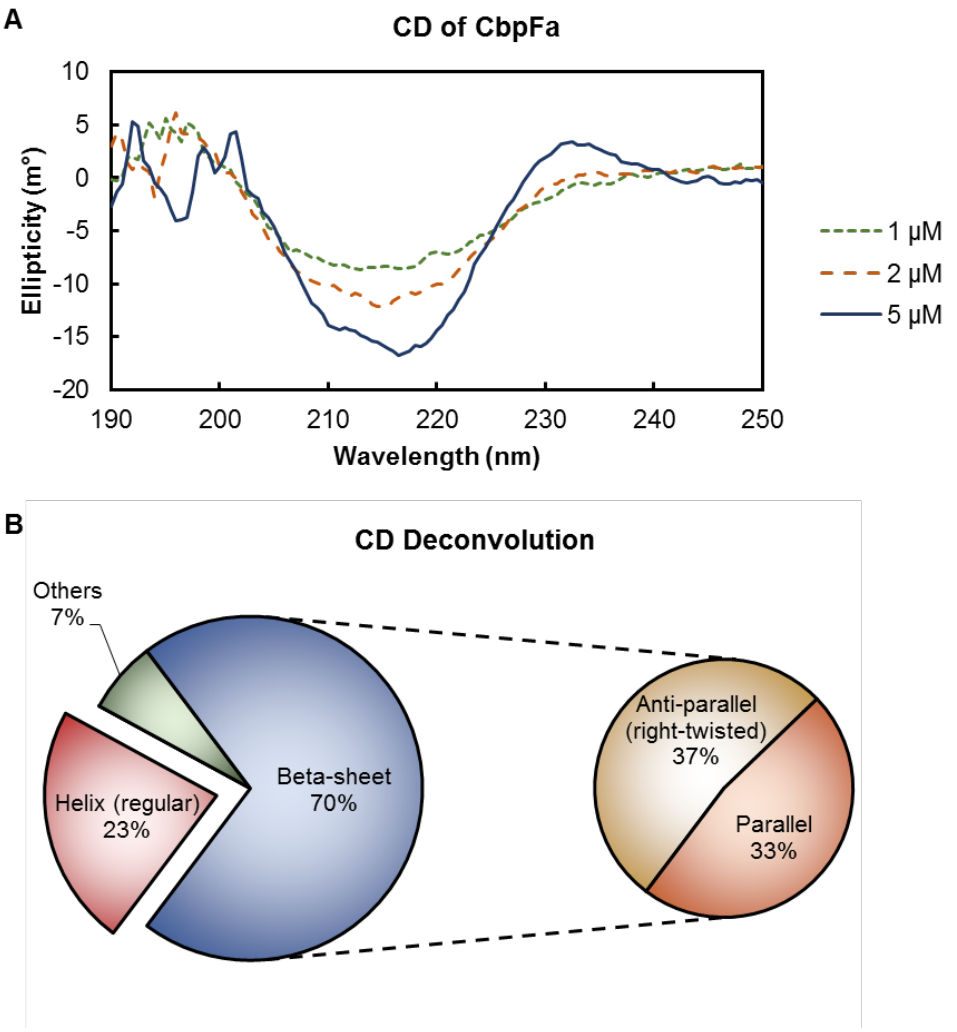


Figure 5.9 | **CD spectra and secondary structure estimations from CbpFa.**

A) Data was collected with wavelengths from 190 to 250 nm (0.5 nm interval) over three concentrations of 1, 2 and 5 µM of recombinant CbpFa 22-330 in SEC Buffer A (**TABLE S 1**). **B)** The CD data was put through a deconvolution program to estimate the relative secondary structure composition (214, 215). The beta-sheet internal compositions will not be as accurate, due to higher background at the shorter wavelengths from with absorbance from chloride ions, however, the ratio of helical to beta-sheet regions is as expected.

5.3.3 Small Angle X-ray Scattering of CbpFa

To assess the overall topology of CbpFa, small-angle X-ray scattering (SAXS) data were collected from protein in solution. Protein from the *E. coli* Rosetta™ 2 (DE3) pLacI::pCFR1 construct (**TABLE 2.3**) was purified using the native large-scale preparation technique (see **SECTION 2.4.2**) and concentrated to 10 mg·ml^{-1} in SEC Buffer B (**TABLE S 1**). 45 µl of protein sample was loaded onto a Superdex 200 Increase (GE Healthcare) column with the eluate connected in-line with the X-ray beam and detector (B21 DLS). Prior to protein loading, the column was equilibrated with 10 column volumes of SEC Buffer B and a 10 mg·ml^{-1} BSA control sample was then run through column and analysed to confirm validity of data collection. X-ray scattering data frames were collected at a rate of 3 frames·sec^{-1} with a sample flow rate of 0.075 ml·min^{-1}. **FIGURE 5.10** shows the HPLC trace for CbpFa at 10 mg·ml^{-1} as well as estimates for the radius of gyration (R_g) for each frame collected. Frames with a consistent R_g following any aggregated protein were then combined and averaged prior to analysis.

Following frame collation and averaging, Guinier fitting was performed using ScÅtter (version 3.1) and a pair-distance, $P(r)$, distribution function was calculated. The averaged data are summarised in **FIGURE 5.11**. One feature that was immediately evident was that the primary scattering peak on the dimensionless Kratky plot greatly deviates from the ideal globular Guinier-Kratky point, which is indicative of a non-globular protein, such as TAA, which are rod-shaped. There are also signs of some disordered regions indicated by the high tail on the plot (**FIGURE 5.11-C**). The comparison of reciprocal-space values, derived from Guinier analysis, and real-space values, derived from the $P(r)$ distribution, is shown in **FIGURE 5.12**. The maximum dimension, d_{max}, was set to 229 Å with a resolution limit, q_{max}, equal to 0.185 Å^{-1}. A summary of the processed data is detailed in **TABLE 5.3**.

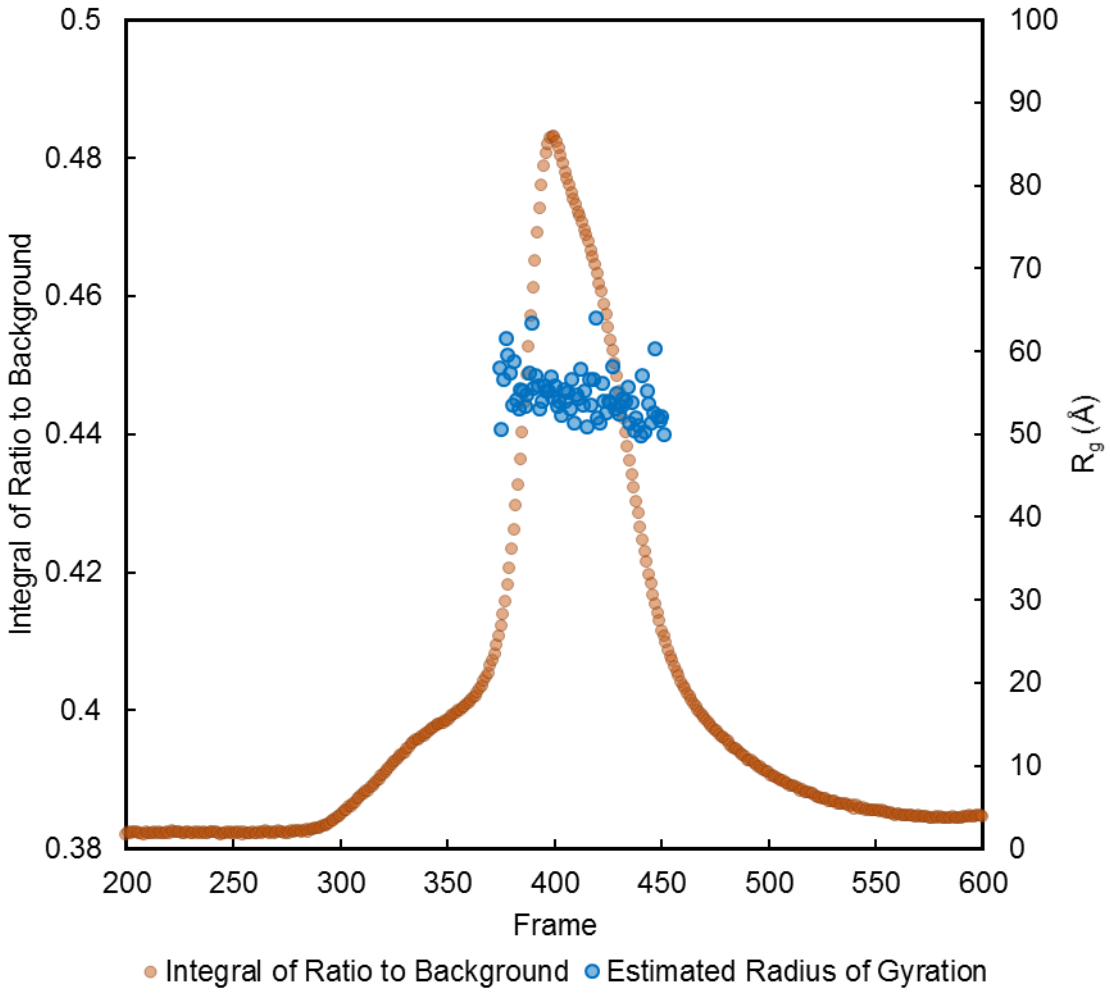


Figure 5.10 | **HPLC Trace for CbpFa with R_g estimates.**

This plot shows the $A_{280\,nm}$ HPLC trace (orange) of the purified protein while it was traversing the X-ray beam. Each point represents one frame of data. Also plotted are the radius of gyration (R_g) estimates (blue). The region prior to the consistent R_g frames had very large R_g values (not shown) and were aggregated protein, hence the earlier elution point. The consistent R_g frames were combined and averaged.

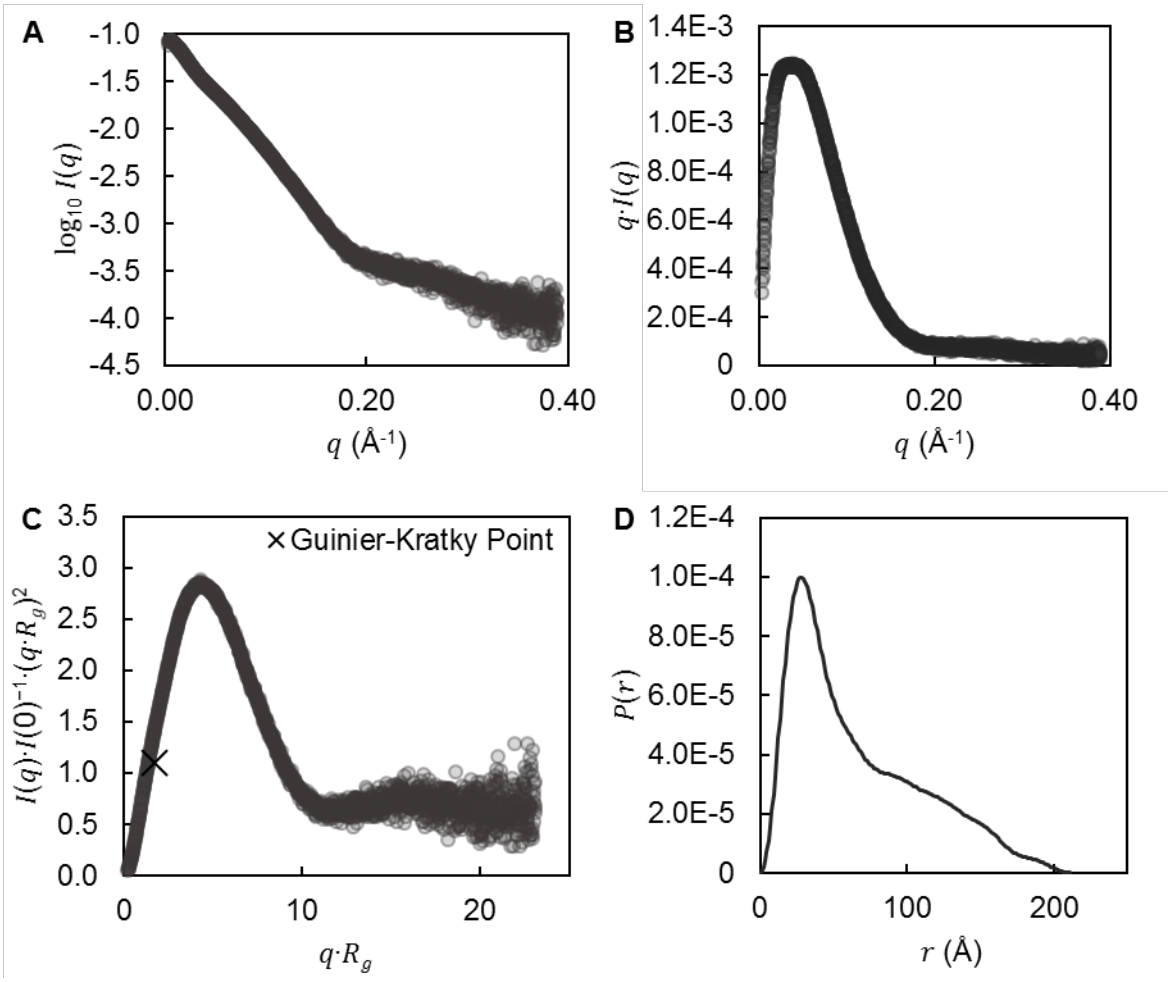


Figure 5.11 | **Summary of SAXS data for CbpFa.**

A) Log_{10} SAXS intensity, $I(q)$, plotted against the scattering vector, q. **B)** Total scattered intensity plot. **C)** Dimensionless Kratky plot with the Guinier-Kratky point displayed $(1.1, \sqrt{3})$, which indicates the primary peak position for globular proteins. As can be seen the main peak for CbpFa is far from this point due to it not being globular and instead, rod-shaped. **D)** Pair-distance, $P(r)$, distribution function for CbpFa with a maximum dimension, d_{max}, set to 229 Å. The fit to the data of the $P(r)$ function is displayed in **FIGURE 5.12**.

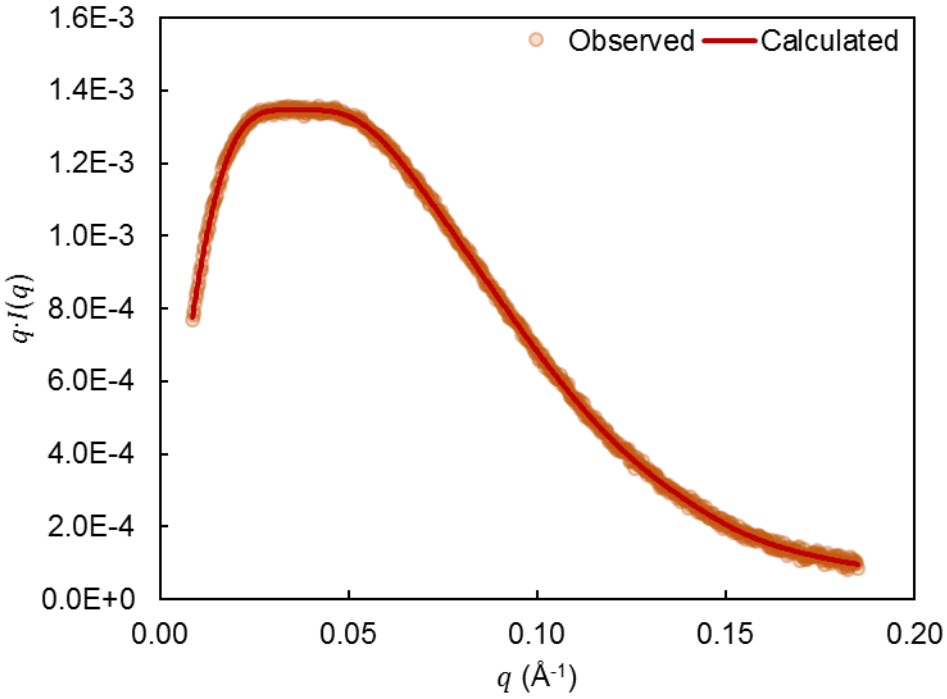


Figure 5.12 | **Observed compared to calculated SAXS data.**

This plot compares the scattering vector, q, against $q \cdot I_{obs}(q)$ (orange points) and $q \cdot I_{calc}(q)$ (red line) obtained from the fitted $P(r)$ function. The limit of resolution, q_{max}, is 0.185 Å^{-1} with a maximum dimension set at 229 Å.

Table 5.3 | **CbpFa SAXS data analysis summary.**

This table lists all the metrics obtained from Guinier analysis and the $P(r)$ distribution and comparisons between the two where applicable. [1] Real space refers to values derived from the $P(r)$ distribution. [2] Reciprocal space refers to values derived from Guinier Analysis.

Metric	**Real Space (R)[1]**	**Reciprocal Space (G)[2]**	**Percentage Difference (%)**	**R_{Error} (+/-)**	**G_{Error} (+/-)**
$I(0)$	8.59 x 10^{-2}	9.55 x 10^{-2}	10.6	6.11 x 10^{-4}	2.29 x 10^{-4}
R_g (Å)	56.8	56.7	1.3	0.79	1.29
Volume	1.47 x 10^{5}	1.50 x 10^{5}	2.0		
M_r (kDa)	88.09	97.96	10.1		
Porod Exponent		3.94			
r (Å)	66.7				
d_{max} (Å)	229				

Following analysis of SAXS data, dummy atom modelling was then performed to approximate the topology of CbpFa. This was achieved by running multiple instances of DAMMIF followed by DAMAVER (216, 217) on the output file from the $P(r)$ distribution. The envelope resolution calculated by SASRES (218) was 37 ± 3 Å. The resulting envelope is shown in **FIGURE 5.13** together with a comparison to the structure of the closest homologue with a known structure, UspA1. The major axis has a maximum length of 230 Å with a maximum diameter of 43 Å on the minor axis. The latter is very similar to that of UspA1 which has a maximum diameter of 45 Å across its minor axis.

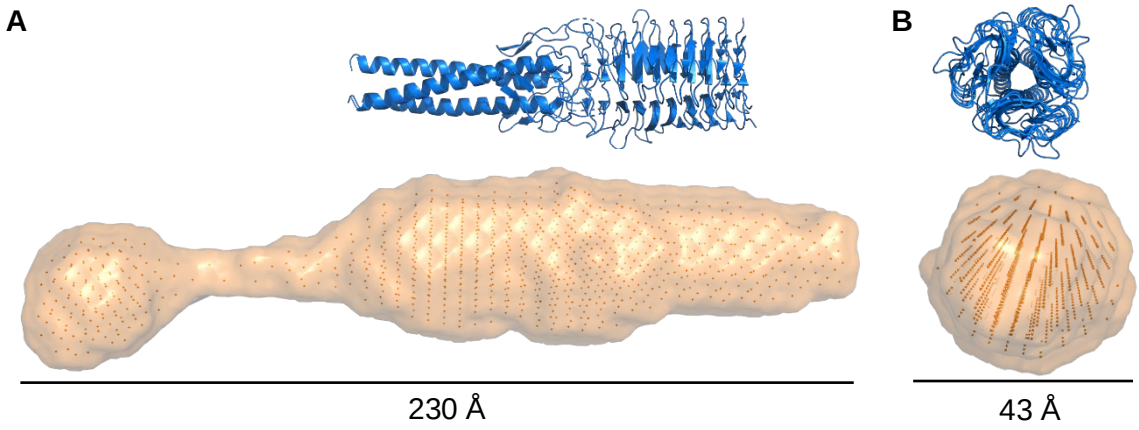


Figure 5.13 | **CbpFa SAXS envelope.**

Shown is the envelope produced following dummy atom modelling with DAMMIN and DAMAVER for CbpFa (orange). **A)** The length of the model was calculated to be 230 Å and shown next the envelope is UspA1 (blue; PDB ID: 3NTN) (102) for comparison. **B)** The cross-section of the envelope had a maximum diameter of 43 Å which is very comparable to UspA1, which has maximum diameter of 45 Å. The UspA1 model used has length of 210 amino residues and CbpFa has 317. The resolution for the envelope was calculated to 37 ± 3 Å using SASRES (218).

5.4 Discussion

From examining the sequences of CbpFs and performing structural predictions, several novel features could be identified. Firstly, it appears that the YadA-like head domains within CbpFs represent a highly conserved domain with an easily identifiable sequence motif (**FIGURE 5.2**). This is in contrast to the head domains of other TAAs that share this similar fold, where there is far less internal conservation of the sequence. The much higher protein sequence conservation found in CbpFs could have nothing to do with protein structure or function but rather as a consequence of *Fusobacterium* having a very low GC % content, therefore having a more restricted codon pool. The GC % content for the YadA-like heads of CbpFa is 34.7 %, about 8 % higher than the global content of *Fn* (27 %). When examining the codon usage for the head domains of CbpFa, AT skewed codons are used preferentially when possible, such as with the case for the central phenylalanine in register position 11 using the TTT codon exclusively. The reason for the elevated content compared to the background primarily comes from the codons encoding the alanine and glycine residues at registers 10 and 12 respectively, as both of these amino acids require codons containing at least two thirds G or C. This points to these two residues being particularly important in maintaining the structure of this domain, otherwise they would have been switched out for lower GC-codon residues such as valine or serine.

Another interesting feature about the YadA-like head domains of CbpFs is their extremely high propensity toward having strongly hydrophobic core consisting of only phenylalanine in CbpFa with the infrequent appearance of leucine, isoleucine or valine across the whole spectrum of CbpFs. This is in opposition to proteins like UspA1, which have a much less conserved motif and more variation at the core register position. For CbpFs, this creates a characteristic ‘phenylalanine zipper’ motif that likely is responsible for the incredible stability of these proteins.

Solving the structure of the stalk domain of these proteins would prove very insightful, as seen in **FIGURE 5.1** and **FIGURE 5.3**, the classical coiled-coil domain cannot be identified from

sequence analysis alone, therefore if one can be found by solving the structure, that could vastly improve prediction software by uncovering a potential novel CC-encoding sequence. For the CbpF a and b proteins, at least, there is likely an insert within the CC-containing domain where there is a glycine-rich sequence followed by a proline (**FIGURE 5.4**) that would break any alpha-helical domain. This is not the case for CbpFc, however, where this domain has been lost.

In addition, to sequence analysis and structure prediction, experimental methods confirm the expected overall predicted topology. Using CD and the latest deconvolution methods (214, 215), the helical to beta-sheet proportions are almost exactly what is expected for CbpFa (**FIGURE 5.9**), though CD deconvolution should always be treated sceptically. Moreover, SAXS analysis on CbpFa, yields an envelope (**FIGURE 5.13)** akin to the expected dimensions. The final piece in the puzzle remains solving the atomic structure. As, single domain expression proved very difficult (**CHAPTER 4**), other methods could be explored such as using GCN4 adaptors to restrict the ends of the expressed domains. This method was used to solve the structure of the TAA SadA from *S. enterica* by solving sequential partial domains fused to GCN4 adaptors (108). Alternatively, as CbpFb crystallised and diffracted well, experimental phasing could yield a solvable solution where molecular replacement could not, such as using iodide phasing as there are too few methionine residues for practicable Se-Met replacement. The structure of the TAA BpaA was solved using iodide phasing (107). Nevertheless, the structures or at least partial structures of the extracellular domains CbpFa and b will hopefully be solved within the next few years, which will expand the current domain dictionary for TAAs (98).

Chapter 6: Discussion

Fusobacterium is a genus of bacteria that has become of recent interest in the scientific community due to its links to colorectal cancer (33-40). These links have yet to be fully established and clarified as there is still no definitive evidence for direct causation and remains a topic of discussion (219, 220). However, there are well-established associations between *Fusobacterium* spp. and pregnancy complications as well as many other human diseases such as IBD and Lemierre's disease for example (49). Moreover, footrot caused by *Fnec* is very well-documented and causes a potentially larger problem from an economic perspective and therefore is where current vaccine development is focussed (68, 69). Nevertheless, the more information that can be unveiled with respect to *Fusobacterium* spp. pathogenesis, would benefit the community.

Historically, the *Fusobacterium* genus consisted of two principal species, *F. necrophorum* and *F. nucleatum*, which were each subdivided into several distinct subspecies. The subspecies for *Fn* were as follows: *Fn* subsp. *nucleatum*, *animalis*, *polymorphum* and *vincentii*. When examining the genomes of the genus, it was found that the subspecies of *Fn* were in fact more distant than originally thought. Other studies also confirmed this discrepancy (22, 23), whereby each of these subspecies should be classified as a unique species. The two subspecies of *Fnec* (*necrophorum* and *funduliforme*), however, are sufficiently related to be considered the same species. Using this new set of criteria based on genomic distance, in this study, two novel species were uncovered when examining previously uncharacterised *Fusobacterium* clinical isolates (**CHAPTER 3**). These species were named *F. oralis* sp. nov. (*For*) and *F. ovarium* sp. nov. (*Fov*). Interestingly, *Fov* was the first documented species of its kind of, which from a collection of only 25 isolates, begs the question of how many more new species of *Fusobacterium* may exist. Furthermore, five *For* strains were within these isolates, therefore, approximately one fifth of all strains examined in this study were unique uncharacterised species. However, an *For* species genome did already exist in the GenBank database (182), but had not been characterised or labelled as

a specific species until this point. In conjunction with the recent identification of the *F. hwasookii* species (25), this suggests that there may still be yet uncharacterised species to be discovered and classified.

From examining the genome sequences of *For*, it was shown they are most closely related to *F. periodonticum* and therefore likely share similar biological characteristics. Moreover, all the *For* isolates examined were from an oral origin which correlates to the normal isolation site for *Fperio* strains. However, there are some notable differences between the two species, most notably the ability to adhere to human CEACAM1 and CEA via a trimeric autotransporter adhesin (CbpF). This could give these strains increased pathogenic strategies when compared to *Fperio*.

Part of the reason that many *Fusobacterium* species have been misclassified is because the current genus is extremely polarised where the clades of *Fn* (and historical subspecies) and *Fnec* are far apart with another small cluster of species branching off approximately halfway in between the two. When examining genomic similarity within the *Fusobacterium* genus, as a control, other non-genus bacteria, such as *Cetobacterium* and *Psychrilyobacter* from the *Fusobacteriaceae* family, were included. When examining species clustering and inter-species difference (**FIGURE 3.7**), it was found that *Fnec* and *F. gonidiaformans* clustered on the level of other genera such as *Ilyobacter*, thus should not exist within the same genus of *F. nucleatum*, and associated species. This assertion is echoed by the very recent bacterial taxonomic reclassification study that normalised taxonomic boundaries based on whole-genome similarity (23). Additionally, the group of species that exist between the two major groups, containing *F. varium*, *F. ulcerans* and *F. mortiferum*, should be split off into their own genus as well. The full list of reclassifications and the respective genus group they belong to are listed in **TABLE S 2**. Combined with the new sequencing data obtained in this research, a total of 188 classifications were made (this equates to 17766 pairwise comparisons per method used).

As discussed earlier in **CHAPTER 3**, there were a number of fringe cases, where certain pairwise genome comparisons did not yield a clear solution. For example, two species of *F. periodonticum* were within the boundary to other species of most *Fperio* strains, however, when compared against each other were below the threshold. This is a notable feature of this particular species, where there is a large amount of heterogeneity making species definitions difficult to draw. The lowest intraspecies Average Nucleotide Identity score was 94.22 %, almost a percent below the acknowledged minimum species boundary of 95 % (180). This represents an interesting anomaly that shows the beginning of evolutionary divergence prior to speciation.

The primary goal of the research conducted in this study was to characterise *Fusobacterium*-CEACAM interactions. These interactions were previously only thought to exist within the *Fn* and *Fv* species (as identified by unpublished work). Therefore, when handling the *For* 2B3 strain, prior to WGS, it was under the assumption that it was an *Fn* or at least *Fn*-like species, such as *Fv*. As this was revealed to be to be an incorrect classification, this eluded to the CbpF gene being on a possible mobile genetic element as other more closely related species, such as *Fpoly*, *Fh* and *Fperio*, did not harbour an analogous gene. This information also provided an improved insight into the evolutionary history of the CbpF proteins, as similar proteins also appeared in *Fa* (though not universally) and *Fov*. The encoding gene was found to exist juxtaposed to the *nik*-operon in all cases where it could be identified. Evidence for this whole region being a mobile element was increased when the genes for the *nik*-operon could not be identified in any strain that did not also harbour the *cbpF* gene indicating they moved together. Moreover, examining the surrounding sequence more closely, a transposase gene could be found further downstream of the *nik*-operon suggesting it was once self-mobilising, however, it is unclear whether it is still mobile.

While all known and putative *cbpF* genes exist within an equivalent genomic surrounding, the resulting CbpF protein sequence varies between species. It appears that the majority of

CbpFs fall into one of two groups: CbpFa (*Fn* and *Fv*) and CbpFb (*For* and some *Fa*), with an isolated case for a third group, CbpFc (*Fov*). This can be seen in **FIGURE 4.6**. The main differences between the two groups is within the variation in the N-terminal head group domain, however, CbpFc also has a deletion in the stalk region that is common to all other CbpFs.

From examining the sequences from all identified CbpF proteins, we were able to identify a highly conserved 14-residue β-roll motif that encodes a YadA-like head fold. CbpF proteins contain a minimum of six repeats of this domain in the N-terminal head group. This cluster of YadA-like head domains underwent expansion at one point in time leading to a discontinuity in the number of repeats of the head group within CbpFs, where CbpFb-like proteins contain around 12 repeats and CbpFa-like have approximately half as many. Interestingly, the final YadA-like head domain (of the initial group) always share the trait of having a tyrosine in the first register position (always an asparagine in other cases; **FIGURE 5.2**; **FIGURE S 6**), indicating the expansion went toward the N-terminus. This whole YadA-like head-containing region likely serves as the N-terminal trimer-stabilising region mediated through a hydrophobic core consisting of primarily of phenylalanine, and to a lesser extent leucine and isoleucine, residues in register position 11 (**FIGURE 5.2**). This region was the most promising for molecular replacement as it shares relatively high homology with the UspA1 protein, though an accurate model could not be obtained.

The most insightful region, if ever solved will be the stalk region, as this perplexes all sequence-analysis and protein modelling software. This region is usually occupied by a trimeric coiled-coil in TAAs, however for the majority of this region, barring the extreme C-terminus directly before the membrane anchor, is obstinately predicted to not contain any CC propensity (using MARCOIL and PCOILS; **FIGURE 5.3**). This region, or at least part of this region, is thought to be responsible for mediating CEACAM1 and CEA adhesion using an analogous mechanism to UspA1 (135). The structure of this region on CbpF proteins should enhance the overall understanding of protein folding and improve prediction

software, as through exhaustive trialling and educated modelling, no definitive complete structure could be solved using molecular replacement on an otherwise promising X-ray diffraction dataset. Short of solving the structure, this study was able to build several promising partial models of the head domain (**FIGURE 5.8**) that could be used later to aid solving the whole structure. Heavy atom phasing will most likely be required to determine the structure of the stalk domain. If a successful structure can be solved for this region, following studies would then likely be able to model the CbpF-CEACAM interactions through computational docking and MD, or even experimentally through X-ray crystallography or SAXS as done for HopQ (138) and UspA1 (135) respectively.

As, identified in **CHAPTER 4**, the interactions between CbpFs and CEACAMs is highly specific with evidence only shown for CEACAM1 and CEA interactions. Additionally, single point mutations made within the probable binding surface on the IgV-like domain of CEACAM1 (CFG face) almost always exhibited diminished adhesion (**FIGURE 4.16**). As well as not showing any specific interactions with other CEACAMs, this shows that the CbpF proteins are highly evolved and therefore probably use a conserved domain facilitate adhesion. By examining confirmed and putative CbpF proteins as well as other TAAs that do not bind CEACAMs, the possible locations for this domain can be greatly reduced to a small region in the stalk domain, though this would need to be confirmed experimentally. As expressing the domain by itself did not yield fruitful results, perhaps using a similar method that constrained regions of the SadA protein using GNC4 adapters (108, 110) may prove to be a more promising route to answering this question. Further studies will also need to characterise any differences in the CEACAM binding profile of CbpFc as currently, no recombinant protein has been produced, so it is unknown whether it can bind other CEACAMs as well as CEACAM1. Additionally, other variants of CbpFa- and CbpFb-like proteins should be examined, such as the CbpFb-like proteins found in a small subset of *Fa* strains.

As well as further studying CbpF-CEACAM interactions, other potential ligands should be considered, such as extracellular matrix components, for example fibronectin or vitronectin. It is known that other TAAs can bind these components, for example both UspA1 and BadA have been shown to bind fibronectin (113, 115).

The identification of the CbpF proteins also sheds light on the difference between certain species and why they may have tendencies to cause different diseases. Many studies have shown pathology delineation between the historical *Fn* subspecies (now distinct species) and highlighted the importance for identifying the specific subspecies (18, 56, 221, 222). For example, *Fa* is the predominant species isolated from intrauterine infections (56). Though other complications arise from the previous incorrect classification of some strains within these species, for example, one labelled *Fpoly* strain in the database, turned out to be an *Fa* species, thus previous studies not including whole-genome sequencing data should be treated with caution. Nonetheless, the apparent divide between the previous *Fn* subspecies could be due to the ability to adhere to CEACAM1 or CEA as these receptors are not universally expressed throughout the body. Elevated expression of both these proteins is seen in the colon for example (119).

Currently, no CbpF-like protein has been identified in any *Fpoly* species or in the majority of *Fa*, however, some strains in this study, as well as at least two other strains in the database, harbour a CbpFb-like gene. This could indicate a reason why some species are more strongly associated with certain pathologies. To confirm the absolute importance of this protein, a study will need to be undertaken to detect the presence of this gene, as well as other species-defining genes, in diseased patients, to examine correlation between disease type, severity and outcome, for example, with the presence of CbpF. This would be particularly insightful for *Fusobacterium*-associated IBD, as CEACAM1 is known to play a role in this disease (223). This would give a clearer indication of the purpose for this protein in disease pathogenesis, as unfortunately, current patient data associated with the strains with whole genomes is sparse, in addition to the small total number of species sequenced.

As previously mentioned, the CbpF proteins only represent one of at least three groups of TAAs possessed by *Fusobacterium*, with two others briefly examined in this study. From data presented in **FIGURE 4.11**, it can be seen that one of these other TAAs (FN0471) indiscriminately binds to HeLa cells and very strongly too. The other TAAs tested, also bound above background levels, indicating that these proteins may facilitate interactions with other receptors present on human cells. More work will be needed to determine the human receptors responsible, but nevertheless, it provides more questions regarding mechanisms of pathogenesis and which interactions are important in mediating disease.

To study all of these proteins further, *Fusobacterium* knockout mutants will need to be created to confirm receptor-binding specificity and possible redundancy. Currently, there is no straightforward way to create *Fusobacterium* knockouts, though some techniques have been described previously using sonoporation and a *Fusobacterium*-specific shuttle plasmid vector (6, 224). The greatly reduced transformation efficiencies associated with *Fusobacterium* spp. is likely due to the presence of a restriction modification system (224). These methods should be explored in the future in an effort to knock the adhesin genes out, with re-complementation with the shuttle vector. In addition to further examining CbpFs, the next TAA candidate adhesin to be studied should be the FN0735 gene product, because homologues to this protein can be found almost universally within *Fn*, *Fv*, *Fa*, *Fpoly*, *Fperio*, *Fh*, *For* and *Fov*, with partial hits in other *Fusobacterium* species. Conversely, the FN0471 gene is found much less commonly. As of yet, almost no information is known regarding these two other TAAs found in *Fusobacterium* other than they show an ability to bind HeLa cells. The universal prevalence of TAAs among the entire genus suggests an intrinsic importance of these proteins for survival and environment adaptation, such as host restriction or expansion, such as the case for *Fnec* species that cause disease in both animals and humans.

As mentioned previously, TAAs have been used successfully as vaccine antigens to elicit immune protection, such as the case for NadA for serogroup B *N. meningitidis* (147, 225).

By adopting a similar rational, proteins such as CbpF and other TAAs within *Fusobacterium* spp. could be viable options in the development of novel vaccines. Their ubiquitous spread and low intraspecies variation would allow specific vaccines that could benefit both humans as well as livestock. As this work focussed primarily on *F. nucleatum* and similar species, further work should be conducted on the other clinically important species of *F. necrophorum*. This study identified TAAs in all *Fnec* strains as part of searching for CbpF homologues; though no direct CbpF homologues were identified, other TAAs were, which showed little variation thus presenting a potential antigenic target. Future work should characterise the properties of these other TAAs and determine whether any epitopes of TAAs from either *Fnec* or *Fn* could elicit a protective immune response.

References

1. Shen XJ, Rawls JF, Randall T, Burcal L, Mpande CN, Jenkins N, et al. Molecular characterization of mucosal adherent bacteria and associations with colorectal adenomas. Gut Microbes. 2010;1(3):138-47.
2. Sobhani I, Tap J, Roudot-Thoraval F, Roperch JP, Letulle S, Langella P, et al. Microbial dysbiosis in colorectal cancer (CRC) patients. PLoS One. 2011;6(1):e16393.
3. Sanapareddy N, Legge RM, Jovov B, McCoy A, Burcal L, Araujo-Perez F, et al. Increased rectal microbial richness is associated with the presence of colorectal adenomas in humans. ISME J. 2012;6(10):1858-68.
4. Desvaux M, Khan A, Beatson SA, Scott-Tucker A, Henderson IR. Protein secretion systems in Fusobacterium nucleatum: genomic identification of Type 4 piliation and complete Type V pathways brings new insight into mechanisms of pathogenesis. Biochim Biophys Acta. 2005;1713(2):92-112.
5. Manson McGuire A, Cochrane K, Griggs AD, Haas BJ, Abeel T, Zeng Q, et al. Evolution of invasion in a diverse set of Fusobacterium species. MBio. 2014;5(6):e01864.
6. Han YW, Ikegami A, Rajanna C, Kawsar HI, Zhou Y, Li M, et al. Identification and characterization of a novel adhesin unique to oral fusobacteria. J Bacteriol. 2005;187(15):5330-40.
7. Spaulding EH, Rettger LF. The Fusobacterium Genus: I. Biochemical and Serological Classification. J Bacteriol. 1937;34(5):535-48.
8. Spaulding EH, Rettger LF. The Fusobacterium Genus: II. Some Observations on Growth Requirements and Variation. J Bacteriol. 1937;34(5):549-63.
9. Bolstad AI, Jensen HB, Bakken V. Taxonomy, biology, and periodontal aspects of Fusobacterium nucleatum. Clin Microbiol Rev. 1996;9(1):55-71.
10. Loesche WJ, Gibbons RJ. Amino acid fermentation by Fusobacterium nucleatum. Arch Oral Biol. 1968;13(2):191-202.

11. Shah HN, Gharbia SE, Zhang MI. Measurement of electrical bioimpedance for studying utilization of amino acids and peptides by Porphyromonas gingivalis, Fusobacterium nucleatum, and Treponema denticola. Clin Infect Dis. 1993;16 Suppl 4:S404-7.
12. Dzink JL, Socransky SS. Amino acid utilization by Fusobacterium nucleatum grown in a chemically defined medium. Oral Microbiol Immunol. 1990;5(3):172-4.
13. Gharbia SE, Shah HN. Glucose-Utilization and Growth-Response to Protein Hydrolysates by Fusobacterium Species. Current Microbiology. 1988;17(4):229-34.
14. Robrish SA, Oliver C, Thompson J. Sugar metabolism by fusobacteria: regulation of transport, phosphorylation, and polymer formation by Fusobacterium mortiferum ATCC 25557. Infect Immun. 1991;59(12):4547-54.
15. Robrish SA, Thompson J. Regulation of fructose metabolism and polymer synthesis by Fusobacterium nucleatum ATCC 10953. J Bacteriol. 1990;172(10):5714-23.
16. Shah HN, Gharbia SE. Ecological events in subgingival dental plaque with reference to Bacteroides and Fusobacterium species. Infection. 1989;17(4):264-8.
17. Gharbia SE, Shah HN. Comparison of the amino acid uptake profile of reference and clinical isolates of Fusobacterium nucleatum subspecies. Oral Microbiol Immunol. 1991;6(5):264-9.
18. Gharbia SE, Shah HN, Lawson PA, Haapasalo M. Distribution and frequency of Fusobacterium nucleatum subspecies in the human oral cavity. Oral Microbiol Immunol. 1990;5(6):324-7.
19. Gharbia SE, Shah HN. Fusobacterium nucleatum subsp. fusiforme subsp. nov. and Fusobacterium nucleatum subsp. animalis subsp. nov. as additional subspecies within Fusobacterium nucleatum. Int J Syst Bacteriol. 1992;42(2):296-8.
20. Kook JK, Park SN, Lim YK, Choi MH, Cho E, Kong SW, et al. Fusobacterium nucleatum subsp. fusiforme Gharbia and Shah 1992 is a later synonym of Fusobacterium nucleatum subsp. vincentii Dzink et al. 1990. Curr Microbiol. 2013;66(4):414-7.

21. Ang MY, Dymock D, Tan JL, Thong MH, Tan QK, Wong GJ, et al. Genome Sequence of Fusobacterium nucleatum Strain W1481, a Possible New Subspecies Isolated from a Periodontal Pocket. Genome Announc. 2014;2(1).
22. Kook JK, Park SN, Lim YK, Cho E, Jo E, Roh H, et al. Genome-Based Reclassification of Fusobacterium nucleatum Subspecies at the Species Level. Curr Microbiol. 2017.
23. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil PA, et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol. 2018.
24. Langworth BF. Fusobacterium necrophorum: its characteristics and role as an animal pathogen. Bacteriol Rev. 1977;41(2):373-90.
25. Cho E, Park SN, Lim YK, Shin Y, Paek J, Hwang CH, et al. Fusobacterium hwasookii sp. nov., Isolated from a Human Periodontitis Lesion. Curr Microbiol. 2015;70(2):169-75.
26. Slots J, Potts TV, Mashimo PA. Fusobacterium periodonticum, a new species from the human oral cavity. J Dent Res. 1983;62(9):960-3.
27. Kolenbrander PE. Oral microbial communities: biofilms, interactions, and genetic systems. Annu Rev Microbiol. 2000;54:413-37.
28. Signat B, Roques C, Poulet P, Duffaut D. Fusobacterium nucleatum in periodontal health and disease. Curr Issues Mol Biol. 2011;13(2):25-36.
29. Hendrickson EL, Wang T, Beck DA, Dickinson BC, Wright CJ, R JL, et al. Proteomics of Fusobacterium nucleatum within a model developing oral microbial community. Microbiologyopen. 2014;3(5):729-51.
30. Zijnge V, van Leeuwen MB, Degener JE, Abbas F, Thurnheer T, Gmur R, et al. Oral biofilm architecture on natural teeth. PLoS One. 2010;5(2):e9321.
31. Li X, Kolltveit KM, Tronstad L, Olsen I. Systemic Diseases Caused by Oral Infection. Clin Microbiol Rev. 2000;13(4):547-58.
32. Hajishengallis G. Periodontitis: from microbial immune subversion to systemic inflammation. Nat Rev Immunol. 2015;15(1):30-44.

33. Castellarin M, Warren RL, Freeman JD, Dreolini L, Krzywinski M, Strauss J, et al. Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. Genome Res. 2012;22(2):299-306.

34. McCoy AN, Araujo-Perez F, Azcarate-Peril A, Yeh JJ, Sandler RS, Keku TO. Fusobacterium is associated with colorectal adenomas. PLoS One. 2013;8(1):e53653.

35. Mima K, Sukawa Y, Nishihara R, Qian ZR, Yamauchi M, Inamura K, et al. Fusobacterium nucleatum and T Cells in Colorectal Carcinoma. JAMA Oncol. 2015;1(5):653-61.

36. Ito M, Kanno S, Nosho K, Sukawa Y, Mitsuhashi K, Kurihara H, et al. Association of Fusobacterium nucleatum with clinical and molecular features in colorectal serrated pathway. Int J Cancer. 2015;137(6):1258-68.

37. Li YY, Ge QX, Cao J, Zhou YJ, Du YL, Shen B, et al. Association of Fusobacterium nucleatum infection with colorectal cancer in Chinese patients. World J Gastroenterol. 2016;22(11):3227-33.

38. Yang Y, Weng W, Peng J, Hong L, Yang L, Toiyama Y, et al. Fusobacterium nucleatum Increases Proliferation of Colorectal Cancer Cells and Tumor Development in Mice by Activating Toll-Like Receptor 4 Signaling to Nuclear Factor-kappaB, and Up-regulating Expression of MicroRNA-21. Gastroenterology. 2017;152(4):851-66 e24.

39. Wong SH, Kwong TNY, Chow TC, Luk AKC, Dai RZW, Nakatsu G, et al. Quantitation of faecal Fusobacterium improves faecal immunochemical test in detecting advanced colorectal neoplasia. Gut. 2017;66(8):1441-8.

40. Kostic AD, Gevers D, Pedamallu CS, Michaud M, Duke F, Earl AM, et al. Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. Genome Res. 2012;22(2):292-8.

41. Kostic AD, Chun E, Robertson L, Glickman JN, Gallini CA, Michaud M, et al. Fusobacterium nucleatum potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. Cell Host Microbe. 2013;14(2):207-15.

42. Yan X, Liu L, Li H, Qin H, Sun Z. Clinical significance of Fusobacterium nucleatum, epithelial-mesenchymal transition, and cancer stem cell markers in stage III/IV colorectal cancer patients. Onco Targets Ther. 2017;10:5031-46.
43. Yamaoka Y, Suehiro Y, Hashimoto S, Hoshida T, Fujimoto M, Watanabe M, et al. Fusobacterium nucleatum as a prognostic marker of colorectal cancer in a Japanese population. J Gastroenterol. 2017.
44. Fardini Y, Wang X, Temoin S, Nithianantham S, Lee D, Shoham M, et al. Fusobacterium nucleatum adhesin FadA binds vascular endothelial cadherin and alters endothelial integrity. Mol Microbiol. 2011;82(6):1468-80.
45. Gur C, Ibrahim Y, Isaacson B, Yamin R, Abed J, Gamliel M, et al. Binding of the Fap2 protein of Fusobacterium nucleatum to human inhibitory receptor TIGIT protects tumors from immune cell attack. Immunity. 2015;42(2):344-55.
46. Rubinstein MR, Wang X, Liu W, Hao Y, Cai G, Han YW. Fusobacterium nucleatum promotes colorectal carcinogenesis by modulating E-cadherin/beta-catenin signaling via its FadA adhesin. Cell Host Microbe. 2013;14(2):195-206.
47. Kwong TNY, Wang X, Nakatsu G, Chow TC, Tipoe T, Dai RZW, et al. Association Between Bacteremia From Specific Microbes and Subsequent Diagnosis of Colorectal Cancer. Gastroenterology. 2018;155(2):383-90.e8.
48. Warren RL, Freeman DJ, Pleasance S, Watson P, Moore RA, Cochrane K, et al. Co-occurrence of anaerobic bacteria in colorectal carcinomas. Microbiome. 2013;1(1):16.
49. Han YW. Fusobacterium nucleatum: a commensal-turned pathogen. Curr Opin Microbiol. 2015;23c:141-7.
50. Bashir A, Miskeen AY, Bhat A, Fazili KM, Ganai BA. Fusobacterium nucleatum: an emerging bug in colorectal tumorigenesis. Eur J Cancer Prev. 2015.
51. Mitsuhashi K, Nosho K, Sukawa Y, Matsunaga Y, Ito M, Kurihara H, et al. Association of Fusobacterium species in pancreatic cancer tissues with molecular features and prognosis. Oncotarget. 2015;6(9):7209-20.

52. Binder Gallimidi A, Fischman S, Revach B, Bulvik R, Maliutina A, Rubinstein AM, et al. Periodontal pathogens Porphyromonas gingivalis and Fusobacterium nucleatum promote tumor progression in an oral-specific chemical carcinogenesis model. Oncotarget. 2015;6(26):22613-23.

53. Han YW, Shen T, Chung P, Buhimschi IA, Buhimschi CS. Uncultivated bacteria as etiologic agents of intra-amniotic inflammation leading to preterm birth. J Clin Microbiol. 2009;47(1):38-47.

54. Han YW, Wang X. Mobile microbiome: oral bacteria in extra-oral infections and inflammation. J Dent Res. 2013;92(6):485-91.

55. Han YW. Oral health and adverse pregnancy outcomes - what's next? J Dent Res. 2011;90(3):289-93.

56. Wang X, Buhimschi CS, Temoin S, Bhandari V, Han YW, Buhimschi IA. Comparative microbial analysis of paired amniotic fluid and cord blood from pregnancies complicated by preterm birth and early-onset neonatal sepsis. PLoS One. 2013;8(2):e56131.

57. Gonzales-Marin C, Spratt DA, Allaker RP. Maternal oral origin of Fusobacterium nucleatum in adverse pregnancy outcomes as determined using the 16S-23S rRNA gene intergenic transcribed spacer region. J Med Microbiol. 2013;62(Pt 1):133-44.

58. Han YW, Fardini Y, Chen C, Iacampo KG, Peraino VA, Shamonki JM, et al. Term stillbirth caused by oral Fusobacterium nucleatum. Obstet Gynecol. 2010;115(2 Pt 2):442-5.

59. Ikegami A, Chung P, Han YW. Complementation of the fadA mutation in Fusobacterium nucleatum demonstrates that the surface-exposed adhesin promotes cellular invasion and placental colonization. Infect Immun. 2009;77(7):3075-9.

60. Han YW, Redline RW, Li M, Yin L, Hill GB, McCormick TS. Fusobacterium nucleatum induces premature and term stillbirths in pregnant mice: implication of oral bacteria in preterm birth. Infect Immun. 2004;72(4):2272-9.

61. Riordan T. Human infection with Fusobacterium necrophorum (Necrobacillosis), with a focus on Lemierre's syndrome. Clin Microbiol Rev. 2007;20(4):622-59.

62. Zheng L, Giri B. Gastrointestinal Variant of Lemierre Syndrome: Fusobacterium nucleatum Bacteremia-Associated Hepatic Vein Thrombosis: a Case Report and Literature Review. Am J Ther. 2016;23(3):e933-6.

63. Sinave CP, Hardy GJ, Fardy PW. The Lemierre syndrome: suppurative thrombophlebitis of the internal jugular vein secondary to oropharyngeal infection. Medicine (Baltimore). 1989;68(2):85-94.

64. Jones JW, Riordan T, Morgan MS. Investigation of postanginal sepsis and Lemierre's syndrome in the South West Peninsula. Commun Dis Public Health. 2001;4(4):278-81.

65. Brazier JS. Human infections with Fusobacterium necrophorum. Anaerobe. 2006;12(4):165-72.

66. Roberts DS, Egerton JR. The aetiology and pathogenesis of ovine foot-rot. II. The pathogenic association of Fusiformis nodosus and F. necrophorus. J Comp Pathol. 1969;79(2):217-27.

67. Bennett G, Hickford J, Sedcole R, Zhou H. Dichelobacter nodosus, Fusobacterium necrophorum and the epidemiology of footrot. Anaerobe. 2009;15(4):173-6.

68. Nieuwhof GJ, Bishop SC. Costs of the major endemic diseases of sheep in Great Britain and the potential benefits of reduction in disease impact. Animal Science. 2005;81:23-9.

69. Dhungyel O, Hunter J, Whittington R. Footrot vaccines and vaccination. Vaccine. 2014;32(26):3139-46.

70. Temoin S, Chakaki A, Askari A, El-Halaby A, Fitzgerald S, Marcus RE, et al. Identification of oral bacterial DNA in synovial fluid of patients with arthritis with native and failed prosthetic joints. J Clin Rheumatol. 2012;18(3):117-21.

71. Strauss J, Kaplan GG, Beck PL, Rioux K, Panaccione R, Devinney R, et al. Invasive potential of gut mucosa-derived Fusobacterium nucleatum positively correlates with IBD status of the host. Inflamm Bowel Dis. 2011;17(9):1971-8.

72. Tahara T, Shibata T, Kawamura T, Okubo M, Ichikawa Y, Sumi K, et al. Fusobacterium detected in colonic biopsy and clinicopathological features of ulcerative colitis in Japan. Dig Dis Sci. 2015;60(1):205-10.

73. Figuero E, Sanchez-Beltran M, Cuesta-Frechoso S, Tejerina JM, del Castro JA, Gutierrez JM, et al. Detection of periodontal bacteria in atheromatous plaque by nested polymerase chain reaction. J Periodontol. 2011;82(10):1469-77.

74. Swidsinski A, Dorffel Y, Loening-Baucke V, Theissig F, Ruckert JC, Ismail M, et al. Acute appendicitis is characterised by local invasion with Fusobacterium nucleatum/necrophorum. Gut. 2011;60(1):34-40.

75. Sparks Stein P, Steffen MJ, Smith C, Jicha G, Ebersole JL, Abner E, et al. Serum antibodies to periodontal pathogens are a risk factor for Alzheimer's disease. Alzheimers Dement. 2012;8(3):196-203.

76. Brook I. Fusobacterial infections in children. J Infect. 1994;28(2):155-65.

77. Agarwal A, Kanekar S, Sabat S, Thamburaj K. Metronidazole-Induced Cerebellar Toxicity. Neurol Int. 2016;8(1):6365.

78. Hutcherson JA, Sinclair KM, Belvin BR, Gui Q, Hoffman PS, Lewis JP. Amixicile, a novel strategy for targeting oral anaerobic pathogens. Sci Rep. 2017;7(1):10474.

79. Swidsinski A, Loening-Baucke V, Bengmark S, Scholze J, Doerffel Y. Bacterial biofilm suppression with antibiotics for ulcerative and indeterminate colitis: consequences of aggressive treatment. Arch Med Res. 2008;39(2):198-204.

80. Kinder SA, Holt SC. Localization of the Fusobacterium nucleatum T18 adhesin activity mediating coaggregation with Porphyromonas gingivalis T22. J Bacteriol. 1993;175(3):840-50.

81. Bachrach G, Rosen G, Bellalou M, Naor R, Sela MN. Identification of a Fusobacterium nucleatum 65 kDa serine protease. Oral Microbiol Immunol. 2004;19(3):155-9.

82. Doron L, Coppenhagen-Glazer S, Ibrahim Y, Eini A, Naor R, Rosen G, et al. Identification and characterization of fusolisin, the Fusobacterium nucleatum autotransporter serine protease. PLoS One. 2014;9(10):e111329.

83. Lee HR, Rhyu IC, Kim HD, Jun HK, Min BM, Lee SH, et al. In-vivo-induced antigenic determinants of Fusobacterium nucleatum subsp. nucleatum. Mol Oral Microbiol. 2011;26(2):164-72.

84. Miao L, Liu Y, Li Q, Wang Z, Li H, Zhang G. Screening and sequence analysis of the hemolysin gene of Fusobacterium necrophorum. Anaerobe. 2010;16(4):402-4.

85. Kleivdal H, Benz R, Tommassen J, Jensen HB. Identification of positively charged residues of FomA porin of Fusobacterium nucleatum which are important for pore function. Eur J Biochem. 1999;260(3):818-24.

86. Haake SK, Wang X. Cloning and expression of FomA, the major outer-membrane protein gene from Fusobacterium nucleatum T18. Arch Oral Biol. 1997;42(1):19-24.

87. Kleivdal H, Benz R, Jensen HB. The Fusobacterium nucleatum major outer-membrane protein (FomA) forms trimeric, water-filled channels in lipid bilayer membranes. Eur J Biochem. 1995;233(1):310-6.

88. Coppenhagen-Glazer S, Sol A, Abed J, Naor R, Zhang X, Han YW, et al. Fap2 of Fusobacterium nucleatum is a galactose inhibitable adhesin, involved in coaggregation, cell adhesion and preterm birth. Infect Immun. 2015.

89. Guo L, Shokeen B, He X, Shi W, Lux R. Streptococcus mutans SpaP binds to RadD of Fusobacterium nucleatum ssp. polymorphum. Mol Oral Microbiol. 2017;32(5):355-64.

90. Kaplan CW, Lux R, Haake SK, Shi W. The Fusobacterium nucleatum outer membrane protein RadD is an arginine-inhibitable adhesin required for inter-species adherence and the structured architecture of multispecies biofilm. Mol Microbiol. 2009;71(1):35-47.

91. Lima BP, Shi W, Lux R. Identification and characterization of a novel Fusobacterium nucleatum adhesin involved in physical interaction and biofilm formation with Streptococcus gordonii. Microbiologyopen. 2017;6(3).

92. Cotter SE, Surana NK, St Geme JW, 3rd. Trimeric autotransporters: a distinct subfamily of autotransporter proteins. Trends Microbiol. 2005;13(5):199-205.

93. Linke D, Riess T, Autenrieth IB, Lupas A, Kempf VA. Trimeric autotransporter adhesins: variable structure, common function. Trends Microbiol. 2006;14(6):264-70.

94. Sikdar R, Peterson JH, Anderson DE, Bernstein HD. Folding of a bacterial integral outer membrane protein is initiated in the periplasm. Nat Commun. 2017;8(1):1309.

95. Desvaux M, Parham NJ, Henderson IR. The autotransporter secretion system. Res Microbiol. 2004;155(2):53-60.

96. Meng G, Surana NK, St Geme JW, 3rd, Waksman G. Structure of the outer membrane translocator domain of the Haemophilus influenzae Hia trimeric autotransporter. EMBO J. 2006;25(11):2297-304.

97. Szczesny P, Lupas A. Domain annotation of trimeric autotransporter adhesins--daTAA. Bioinformatics. 2008;24(10):1251-6.

98. Bassler J, Hernandez Alvarez B, Hartmann MD, Lupas AN. A domain dictionary of trimeric autotransporter adhesins. Int J Med Microbiol. 2015;305(2):265-75.

99. Mikula KM, Leo JC, Łyskowski A, Kedracka-Krok S, Pirog A, Goldman A. The translocation domain in trimeric autotransporter adhesins is necessary and sufficient for trimerization and autotransportation. J Bacteriol. 2012;194(4):827-38.

100. Cotter SE, Surana NK, Grass S, St Geme JW, 3rd. Trimeric autotransporters require trimerization of the passenger domain for stability and adhesive activity. J Bacteriol. 2006;188(15):5400-7.

101. Kajava AV, Steven AC. The turn of the screw: variations of the abundant beta-solenoid motif in passenger domains of Type V secretory proteins. J Struct Biol. 2006;155(2):306-15.

102. Agnew C, Borodina E, Zaccai NR, Conners R, Burton NM, Vicary JA, et al. Correlation of in situ mechanosensitive responses of the Moraxella catarrhalis adhesin UspA1 with fibronectin and receptor CEACAM1 binding. Proc Natl Acad Sci U S A. 2011;108(37):15174-8.

103. Nummelin H, Merckel MC, Leo JC, Lankinen H, Skurnik M, Goldman A. The Yersinia adhesin YadA collagen-binding domain structure is a novel left-handed parallel beta-roll. EMBO J. 2004;23(4):701-11.

104. Kaiser PO, Riess T, Wagner CL, Linke D, Lupas AN, Schwarz H, et al. The head of Bartonella adhesin A is crucial for host cell interaction of Bartonella henselae. Cell Microbiol. 2008;10(11):2223-34.

105. Leo JC, Lyskowski A, Hattula K, Hartmann MD, Schwarz H, Butcher SJ, et al. The structure of E. coli IgG-binding protein D suggests a general model for bending and binding in trimeric autotransporter adhesins. Structure. 2011;19(7):1021-30.
106. van Ulsen P, Rahman S, Jong WS, Daleke-Schermerhorn MH, Luirink J. Type V secretion: from biogenesis to biotechnology. Biochim Biophys Acta. 2014;1843(8):1592-611.
107. Edwards TE, Phan I, Abendroth J, Dieterich SH, Masoudi A, Guo W, et al. Structure of a Burkholderia pseudomallei trimeric autotransporter adhesin head. PLoS One. 2010;5(9).
108. Hartmann MD, Grin I, Dunin-Horkawicz S, Deiss S, Linke D, Lupas AN, et al. Complete fiber structures of complex trimeric autotransporter adhesins conserved in enterobacteria. Proc Natl Acad Sci U S A. 2012;109(51):20907-12.
109. Szczesny P, Linke D, Ursinus A, Bar K, Schwarz H, Riess TM, et al. Structure of the head of the Bartonella adhesin BadA. PLoS Pathog. 2008;4(8):e1000119.
110. Hartmann MD, Ridderbusch O, Zeth K, Albrecht R, Testa O, Woolfson DN, et al. A coiled-coil motif that sequesters ions to the hydrophobic core. Proc Natl Acad Sci U S A. 2009;106(40):16950-5.
111. Meng G, St Geme JW, Waksman G. Repetitive architecture of the Haemophilus influenzae Hia trimeric autotransporter. J Mol Biol. 2008;384(4):824-36.
112. Zimmerman SM, Michel F, Hogan RJ, Lafontaine ER. The Autotransporter BpaB Contributes to the Virulence of Burkholderia mallei in an Aerosol Model of Infection. PLoS One. 2015;10(5):e0126437.
113. Kaiser PO, Linke D, Schwarz H, Leo JC, Kempf VA. Analysis of the BadA stalk from Bartonella henselae reveals domain-specific and domain-overlapping functions in the host cell infection process. Cell Microbiol. 2012;14(2):198-209.
114. Hill DJ, Virji M. A novel cell-binding mechanism of Moraxella catarrhalis ubiquitous surface protein UspA: specific targeting of the N-domain of carcinoembryonic antigen-related cell adhesion molecules by UspA1. Mol Microbiol. 2003;48(1):117-29.

115. Tan TT, Nordstrom T, Forsgren A, Riesbeck K. The respiratory pathogen Moraxella catarrhalis adheres to epithelial cells by interacting with fibronectin through ubiquitous surface proteins A1 and A2. J Infect Dis. 2005;192(6):1029-38.

116. Tan TT, Forsgren A, Riesbeck K. The respiratory pathogen moraxella catarrhalis binds to laminin via ubiquitous surface proteins A1 and A2. J Infect Dis. 2006;194(4):493-7.

117. Attia AS, Ram S, Rice PA, Hansen EJ. Binding of vitronectin by the Moraxella catarrhalis UspA2 protein interferes with late stages of the complement cascade. Infect Immun. 2006;74(3):1597-611.

118. Hill DJ, Edwards AM, Rowe HA, Virji M. Carcinoembryonic antigen-related cell adhesion molecule (CEACAM)-binding recombinant polypeptide confers protection against infection by respiratory and urogenital pathogens. Mol Microbiol. 2005;55(5):1515-27.

119. Fagerberg L, Hallström BM, Oksvold P, Kampf C, Djureinovic D, Odeberg J, et al. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. Mol Cell Proteomics. 2014;13(2):397-406.

120. Beauchemin N, Draber P, Dveksler G, Gold P, Gray-Owen S, Grunert F, et al. Redefined nomenclature for members of the carcinoembryonic antigen family. Exp Cell Res. 1999;252(2):243-9.

121. Kammerer R, Zimmermann W. Coevolution of activating and inhibitory receptors within mammalian carcinoembryonic antigen families. BMC Biol. 2010;8:12.

122. Kuespert K, Pils S, Hauck CR. CEACAMs: their role in physiology and pathophysiology. Curr Opin Cell Biol. 18. United States2006. p. 565-71.

123. Ocklind C, Forsum U, Obrink B. Cell surface localization and tissue distribution of a hepatocyte cell-cell adhesion glycoprotein (cell-CAM 105). J Cell Biol. 1983;96(4):1168-71.

124. Rojas M, Fuks A, Stanners CP. Biliary glycoprotein, a member of the immunoglobulin supergene family, functions in vitro as a Ca2(+)-dependent intercellular adhesion molecule. Cell Growth Differ. 1990;1(11):527-33.

125. Oikawa S, Inuzuka C, Kuroki M, Matsuoka Y, Kosaki G, Nakazato H. Cell adhesion activity of non-specific cross-reacting antigen (NCA) and carcinoembryonic antigen (CEA)

expressed on CHO cell surface: homophilic and heterophilic adhesion. Biochem Biophys Res Commun. 1989;164(1):39-45.

126. Oikawa S, Inuzuka C, Kuroki M, Arakawa F, Matsuoka Y, Kosaki G, et al. A specific heterotypic cell adhesion activity between members of carcinoembryonic antigen family, W272 and NCA, is mediated by N-domains. J Biol Chem. 1991;266(13):7995-8001.
127. Chen T, Zimmermann W, Parker J, Chen I, Maeda A, Bolland S. Biliary glycoprotein (BGPa, CD66a, CEACAM1) mediates inhibitory signals. J Leukoc Biol. 2001;70(2):335-40.
128. McCaw SE, Schneider J, Liao EH, Zimmermann W, Gray-Owen SD. Immunoreceptor tyrosine-based activation motif phosphorylation during engulfment of Neisseria gonorrhoeae by the neutrophil-restricted CEACAM3 (CD66d) receptor. Mol Microbiol. 2003;49(3):623-37.
129. Hau J, Gidley-Baird AA, Westergaard JG, Teisner B. The effect on pregnancy of intrauterine administration of antibodies against two pregnancy-associated murine proteins: murine pregnancy-specific beta 1-glycoprotein and murine pregnancy-associated alpha 2-glycoprotein. Biomed Biochim Acta. 1985;44(7-8):1255-9.
130. Ergün S, Kilik N, Ziegeler G, Hansen A, Nollau P, Götze J, et al. CEA-related cell adhesion molecule 1: a potent angiogenic factor and a major effector of vascular endothelial growth factor. Mol Cell. 2000;5(2):311-20.
131. Markel G, Achdout H, Katz G, Ling KL, Salio M, Gruda R, et al. Biological function of the soluble CEACAM1 protein and implications in TAP2-deficient patients. Eur J Immunol. 2004;34(8):2138-48.
132. Barclay AN. Membrane proteins with immunoglobulin-like domains--a master superfamily of interaction molecules. Semin Immunol. 2003;15(4):215-23.
133. Virji M, Evans D, Hadfield A, Grunert F, Teixeira AM, Watt SM. Critical determinants of host receptor targeting by Neisseria meningitidis and Neisseria gonorrhoeae: identification of Opa adhesiotopes on the N-domain of CD66 molecules. Mol Microbiol. 1999;34(3):538-51.

134. Heinrich A, Heyl KA, Klaile E, Müller MM, Klassert TE, Wiessner A, et al. Moraxella catarrhalis induces CEACAM3-Syk-CARD9-dependent activation of human granulocytes. Cell Microbiol. 2016;18(11):1570-82.

135. Conners R, Hill DJ, Borodina E, Agnew C, Daniell SJ, Burton NM, et al. The Moraxella adhesin UspA1 binds to its human CEACAM1 receptor by a deformable trimeric coiled-coil. EMBO J. 2008;27(12):1779-89.

136. Xie Q, Brackenbury LS, Hill DJ, Williams NA, Qu X, Virji M. Moraxella catarrhalis Adhesin UspA1-derived Recombinant Fragment rD-7 Induces Monocyte Differentiation to CD14+CD206+ Phenotype. PLoS One. 92014.

137. Bonsor DA, Günther S, Beadenkopf R, Beckett D, Sundberg EJ. Diverse oligomeric states of CEACAM IgV domains. Proc Natl Acad Sci U S A. 2015;112(44):13561-6.

138. Moonens K, Hamway Y, Neddermann M, Reschke M, Tegtmeyer N, Kruse T, et al. H*elicobacter pylori* adhesin HopQ disrupts t*rans* dimerization in human CEACAMs. EMBO J. 2018;37(13).

139. Korotkova N, Yang Y, Le Trong I, Cota E, Demeler B, Marchant J, et al. Binding of Dr adhesins of Escherichia coli to carcinoembryonic antigen triggers receptor dissociation. Mol Microbiol. 2008;67(2):420-34.

140. Königer V, Holsten L, Harrison U, Busch B, Loell E, Zhao Q, et al. Erratum: Helicobacter pylori exploits human CEACAMs via HopQ for adherence and translocation of CagA. Nat Microbiol. 2016;2:16233.

141. Kammerer R, Rüttiger L, Riesenberg R, Schäuble C, Krupar R, Kamp A, et al. Loss of mammal-specific tectorial membrane component carcinoembryonic antigen cell adhesion molecule 16 (CEACAM16) leads to hearing impairment at low and high frequencies. J Biol Chem. 2012;287(26):21584-98.

142. Cheatham MA, Goodyear RJ, Homma K, Legan PK, Korchagina J, Naskar S, et al. Loss of the tectorial membrane protein CEACAM16 enhances spontaneous, stimulus-frequency, and transiently evoked otoacoustic emissions. J Neurosci. 2014;34(31):10325-38.

143. Hofrichter MA, Nanda I, Gräf J, Schröder J, Shehata-Dieler W, Vona B, et al. A Novel de novo Mutation in CEACAM16 Associated with Postlingual Hearing Impairment. Mol Syndromol. 2015;6(4):156-63.
144. Wang H, Wang X, He C, Li H, Qing J, Grati M, et al. Exome sequencing identifies a novel CEACAM16 mutation associated with autosomal dominant nonsyndromic hearing loss DFNA4B in a Chinese family. J Hum Genet. 2015;60(3):119-26.
145. Camacho-Leal P, Stanners CP. The human carcinoembryonic antigen (CEA) GPI anchor mediates anoikis inhibition by inactivation of the intrinsic death pathway. Oncogene. 2007;27(11):1545-53.
146. Gold P, Freedman SO. Specific carcinoembryonic antigens of the human digestive system. J Exp Med. 1965;122(3):467-81.
147. Gorringe AR, Pajón R. Bexsero: a multicomponent vaccine for prevention of meningococcal disease. Hum Vaccin Immunother. 2012;8(2):174-83.
148. Comanducci M, Bambini S, Brunelli B, Adu-Bobie J, Aricò B, Capecchi B, et al. NadA, a novel vaccine candidate of Neisseria meningitidis. J Exp Med. 2002;195(11):1445-54.
149. Berrow NS, Alderton D, Sainsbury S, Nettleship J, Assenberg R, Rahman N, et al. A versatile ligation-independent cloning method suitable for high-throughput expression screening applications. Nucleic Acids Res. 2007;35(6):e45.
150. Wilkins MR, Gasteiger E, Bairoch A, Sanchez JC, Williams KL, Appel RD, et al. Protein identification and analysis tools in the ExPASy server. Methods Mol Biol. 1999;112:531-52.
151. Battye TG, Kontogiannis L, Johnson O, Powell HR, Leslie AG. iMOSFLM: a new graphical interface for diffraction-image processing with MOSFLM. Acta Crystallogr D Biol Crystallogr. 2011;67(Pt 4):271-81.
152. Winn MD, Ballard CC, Cowtan KD, Dodson EJ, Emsley P, Evans PR, et al. Overview of the CCP4 suite and current developments. Acta Crystallogr D Biol Crystallogr. 2011;67(Pt 4):235-42.

153. Adams PD, Afonine PV, Bunkóczi G, Chen VB, Davis IW, Echols N, et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr D Biol Crystallogr. 2010;66(Pt 2):213-21.
154. Berendsen HJC, Vanderspoel D, Vandrunen R. Gromacs - a Message-Passing Parallel Molecular-Dynamics Implementation. Computer Physics Communications. 1995;91(1-3):43-56.
155. Abraham M J MT, Schulz R, Páll S, Smith J C, Hess S B, Lindahl E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX. 2015;Volumes 1–2:19–25.
156. Schmid N, Eichenberger AP, Choutko A, Riniker S, Winger M, Mark AE, et al. Definition and testing of the GROMOS force-field versions 54A7 and 54B7. Eur Biophys J. 2011;40(7):843-56.
157. Kaminski GA, Friesner RA, Tirado-Rives J, Jorgensen WL. Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins via Comparison with Accurate Quantum Chemical Calculations on Peptides. The Journal of Physical Chemistry B. 2001;105(28):6474-87.
158. Berendsen HJC, Grigera JR, Straatsma TP. The missing term in effective pair potentials. The Journal of Physical Chemistry. 1987;91(24):6269-71.
159. GenBank Bacteria Genome Assembly Repository: National Center for Biotechnology Information; [Available from: ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/bacteria/.
160. RefSeq Bacteria Genome Assembly Repository: National Center for Biotechnology Information; [Available from: ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/.
161. Lee I, Kim YO, Park SC, Chun J. OrthoANI: An improved algorithm and software for calculating average nucleotide identity. Int J Syst Evol Microbiol. 2015;66:1100-3.
162. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26(19):2460-1.
163. Deloger M, El Karoui M, Petit MA. A genomic distance based on MUM indicates discontinuity between most bacterial species and genera. J Bacteriol. 2009;191(1):91-9.

164. Brewer ML. MUMmer and MUMi Python API 2018 [Available from: https://github.com/mb1511/MUMmer-MUMi.
165. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.
166. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792-7.
167. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol. 2016;33(7):1870-4.
168. Brewer ML. Pan-locus Sequence Analysis 2018 [Available from: https://github.com/mb1511/PLSA.
169. Huson DH. SplitsTree: analyzing and visualizing evolutionary data. Bioinformatics. 1998;14(1):68-73.
170. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987;4(4):406-25.
171. Gascuel O. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol. 1997;14(7):685-95.
172. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011;7:539.
173. Brewer ML. Genome Hive Plots 2018 [Available from: https://github.com/mb1511/GenomeWeb.
174. R: A language and environment for statistical computing. 2013.
175. Ang MY, Dutta A, Wee WY, Dymock D, Paterson IC, Choo SW. Comparative Genome Analysis of Fusobacterium nucleatum. Genome Biol Evol. 2016.
176. Stackebrandt E, Goebel BM. Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. International Journal of Systematic and Evolutionary Microbiology. 1994;44:846-9.

177. McKay TL, Ko J, Bilalis Y, DiRienzo JM. Mobile genetic elements of Fusobacterium nucleatum. Plasmid. 1995;33(1):15-25.

178. Leung DW, Lui AC, Merilees H, McBride BC, Smith M. A restriction enzyme from Fusobacterium nucleatum 4H which recognizes GCNGC. Nucleic Acids Res. 1979;6(1):17-25.

179. Lui AC, McBride BC, Vovis GF, Smith M. Site specific endonuclease from Fusobacterium nucleatum. Nucleic Acids Res. 1979;6(1):1-15.

180. Richter M, Rossello-Mora R. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A. 2009;106(45):19126-31.

181. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biol. 2004;5(2):R12.

182. Information NCfB. GenBank Bacteria Genome Assembly Repository [Available from: ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/bacteria/.

183. Zuckerkandl E, Pauling L. Molecules as Documents of Evolutionary History. Journal of Theoretical Biology. 1965;8(2):357-+.

184. Dorsch M, Lovet DN, Bailey GD. Fusobacterium equinum sp. nov., from the oral cavity of horses. Int J Syst Evol Microbiol. 2001;51(Pt 6):1959-63.

185. Sekizuka T, Ogasawara Y, Ohkusa T, Kuroda M. Characterization of Fusobacterium varium Fv113-g1 isolated from a patient with ulcerative colitis based on complete genome sequence and transcriptome analysis. PLoS One. 2017;12(12):e0189319.

186. Sikorski J, Chertkov O, Lapidus A, Nolan M, Lucas S, Del Rio TG, et al. Complete genome sequence of Ilyobacter polytropus type strain (CuHbu1). Stand Genomic Sci. 2010;3(3):304-14.

187. Zhao JS, Manno D, Hawari J. Psychrilyobacter atlanticus gen. nov., sp. nov., a marine member of the phylum Fusobacteria that produces H2 and degrades nitramine explosives under low temperature conditions. Int J Syst Evol Microbiol. 2009;59(Pt 3):491-7.

188. Parks DH, Rinke C, Chuvochina M, Chaumeil PA, Woodcroft BJ, Evans PN, et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol. 2017;2(11):1533-42.

189. Finegold SM, Vaisanen ML, Molitoris DR, Tomzynski TJ, Song Y, Liu C, et al. Cetobacterium somerae sp. nov. from human feces and emended description of the genus Cetobacterium. Syst Appl Microbiol. 2003;26(2):177-81.

190. Foster G, Ross HM, Naylor RD, Collins MD, Ramos CP, Fernandez Garayzabal F, et al. Cetobacterium ceti gen. nov., sp. nov., a new gram-negative obligate anaerobe from sea mammals. Lett Appl Microbiol. 1995;21(3):202-6.

191. Kim HS, Lee DS, Chang YH, Kim MJ, Koh S, Kim J, et al. Application of rpoB and zinc protease gene for use in molecular discrimination of Fusobacterium nucleatum subspecies. J Clin Microbiol. 2010;48(2):545-53.

192. Nie S, Tian B, Wang X, Pincus DH, Welker M, Gilhuley K, et al. Fusobacterium nucleatum Subspecies Identification by Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry. J Clin Microbiol. 53. United States: American Society for Microbiology. All Rights Reserved.; 2015. p. 1399-402.

193. Felsenstein J. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. Evolution. 1985;39(4):783-91.

194. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 1970;48(3):443-53.

195. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A. 1992;89(22):10915-9.

196. Tchoupa AK, Schuhmacher T, Hauck CR. Signaling by epithelial members of the CEACAM family – mucosal docking sites for pathogenic bacteria. Cell Commun Signal. 122014. p. 27.

197. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004;14(6):1188-90.

198. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 2018;46(W1):W296-W303.
199. Fernandez-Fuentes N, Madrid-Aliste CJ, Rai BK, Fajardo JE, Fiser A. M4T: a comparative protein structure modeling server. Nucleic Acids Res. 2007;35(Web Server issue):W363-8.
200. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER Suite: protein structure and function prediction. Nat Methods. 2015;12(1):7-8.
201. Xu D, Zhang Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins. 2012;80(7):1715-35.
202. Källberg M, Wang H, Wang S, Peng J, Wang Z, Lu H, et al. Template-based protein structure modeling using the RaptorX web server. Nat Protoc. 2012;7(8):1511-22.
203. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000;28(1):235-42.
204. Wood CW, Woolfson DN. CCBuilder 2.0: Powerful and accessible coiled-coil modeling. Protein Sci. 2018;27(1):103-11.
205. Evans PR, Murshudov GN. How good are my data and what is the resolution? Acta Crystallogr D Biol Crystallogr. 2013;69(Pt 7):1204-14.
206. Evans P. Scaling and assessment of data quality. Acta Crystallogr D Biol Crystallogr. 2006;62(Pt 1):72-82.
207. Evans PR. An introduction to data reduction: space-group determination, scaling and intensity statistics. Acta Crystallogr D Biol Crystallogr. 2011;67(Pt 4):282-92.
208. Bunkóczi G, Read RJ. Improvement of molecular-replacement models with Sculptor. Acta Crystallogr D Biol Crystallogr. 2011;67(Pt 4):303-12.
209. McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, Read RJ. Phaser crystallographic software. J Appl Crystallogr. 2007;40(Pt 4):658-74.
210. Cowtan K. Modified phased translation functions and their application to molecular-fragment location. Acta Crystallogr D Biol Crystallogr. 1998;54(Pt 5):750-6.
211. Sheldrick GM. A short history of SHELX. Acta Crystallogr A. 2008;64(Pt 1):112-22.

212. Emsley P, Lohkamp B, Scott WG, Cowtan K. Features and development of Coot. Acta Crystallogr D Biol Crystallogr. 2010;66(Pt 4):486-501.

213. Vagin AA, Steiner RA, Lebedev AA, Potterton L, McNicholas S, Long F, et al. REFMAC5 dictionary: organization of prior chemical knowledge and guidelines for its use. Acta Crystallogr D Biol Crystallogr. 2004;60(Pt 12 Pt 1):2184-95.

214. Micsonai A, Wien F, Kernya L, Lee YH, Goto Y, Refregiers M, et al. Accurate secondary structure prediction and fold recognition for circular dichroism spectroscopy. Proc Natl Acad Sci U S A. 2015;112(24):E3095-103.

215. Micsonai A, Wien F, Bulyaki E, Kun J, Moussong E, Lee YH, et al. BeStSel: a web server for accurate protein secondary structure prediction and fold recognition from the circular dichroism spectra. Nucleic Acids Res. 2018;46(W1):W315-W22.

216. Franke D, Svergun DI. DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering. J Appl Crystallogr. 2009;42(Pt 2):342-6.

217. Volkov VV, Svergun DI. Uniqueness of ab initio shape determination in small-angle scattering. Journal of Applied Crystallography. 2003;36:860-4.

218. Tuukkanen AT, Kleywegt GJ, Svergun DI. Resolution of ab initio shapes determined from small-angle scattering. IUCrJ. 2016;3(Pt 6):440-7.

219. Keku TO, McCoy AN, Azcarate-Peril AM. Fusobacterium spp. and colorectal cancer: cause or consequence? Trends Microbiol. 2013;21(10):506-8.

220. Amitay EL, Werner S, Vital M, Pieper DH, Hofler D, Gierse IJ, et al. Fusobacterium and colorectal cancer: causal factor or passenger? Results from a large colorectal cancer screening study. Carcinogenesis. 2017;38(8):781-8.

221. Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Costea PI, et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. Mol Syst Biol. 2014;10:766.

222. Ye X, Wang R, Bhattacharya R, Boulbes DR, Fan F, Xia L, et al. Fusobacterium Nucleatum Subspecies Animalis Influences Proinflammatory Cytokine Expression and Monocyte Activation in Human Colorectal Tumors. Cancer Prev Res (Phila). 2017;10(7):398-409.

223. Nagaishi T, Chen Z, Chen L, Iijima H, Nakajima A, Blumberg RS. CEACAM1 and the regulation of mucosal inflammation. Mucosal Immunol. 2008;1 Suppl 1:S39-42.

224. Haake SK, Yoder SC, Attarian G, Podkaminer K. Native Plasmids of Fusobacterium nucleatum: Characterization and Use in Development of Genetic Systems†. J Bacteriol. 1822000. p. 1176-80.

225. Semchenko EA, Tan A, Borrow R, Seib KL. The serogroup B meningococcal vaccine Bexsero elicits antibodies to Neisseria gonorrhoeae. Clin Infect Dis. 2018.

Appendix A: Buffer Compositions

Table S 1 | **Buffer compositions used throughout project.**

Any deviations or additions to these buffers are mentioned within the main text. Unless otherwise stated: pH was adjusted using HCl or NaOH; buffer concentrations are given at final working concentrations; reagents are from Merck Sigma-Aldrich. [1] Thermo Fisher Scientific. [2] For use on the Diamond Light Source B21 SAXS HPLC. [3] Severn Biotech Ltd. [4] Stock solutions of all components were made prior to formulation using a variety of solvents such as methanol, ethanol, DMSO and water.

Buffer Name	Composition	pH
Native Buffer A	50 mM Trizma®-HCl, 200 mM NaCl[1], 20 mM imidazole	7.8
Native Buffer B	50 mM Trizma®-HCl, 200 mM NaCl, 500 mM imidazole	7.8
SEC Buffer A	50 mM Trizma®-HCl, 100 mM NaCl	7.5
SEC Buffer B[2]	100 mM Trizma®-HCl, 100 mM NaCl	7.2
Denaturing Buffer A	50 mM Trizma®-HCl, 200 mM NaCl, 20 mM imidazole, 8 M urea[1]	7.8
Denaturing Buffer B	50 mM Trizma®-HCl, 200 mM NaCl, 500 mM imidazole, 8 M urea	7.8
PAGE-Sample Buffer (4 X)	200 mM Trizma®-HCl, 40 % (v/v) glycerol[1], 8 % (w/v) SDS[1], 5 % (v/v) β-mercaptoehtanol, 0.05 % (w/v) bromophenol blue	6.8
SDS-PAGE Running Buffer	25 mM Trizma®-Base, 200 mM glycine[3], 0.1 % (w/v) SDS	8.5
Transfer Buffer	25 mM Trizma®-Base, 200 mM glycine, 0.1 % (w/v) SDS, 20 % (v/v) methanol	8.5
AP Buffer	1 M Trizma®-HCl, 5 M NaCl, 1 M $MgCl_2$, 0.05% (v/v) TWEEN®-20	9.5
PBS	137 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, 1.8 mM KH_2PO_4	7.4
Carbonate Buffer	50 mM Na_2CO_3, 50 mM $NaHCO_3$	9.6
ELISA Wash Buffer	137 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, 1.8 mM KH_2PO_4, 0.05 % (v/v) TWEEN®-20	7.4
Protein A Loading Buffer	20 mM NaH_2PO_4, 250 mM NaCl	8.0

Protein A Elution Buffer	200 mM Na_2HPO_4, 100 mM citric acid	3.0
Protein A Neutralisation Buffer	1 M Trizma®-HCl, 1 M NaCl	7.5
Protease Inhibitor Cocktail (100 X)[4]	100 mM PMSF, 100 µM E-64, 100 µM Pepstatin A, 6 µM Bestatin, 10 mM EDTA	NA
TBE	90 mM Trizma®-Base, 90 mM boric acid, 2 mM EDTA	8.0

Appendix B: CEACAM Numbering Conventions

Previous studies have labelled residues by mature protein sequence only so, for historical reasons, some proteins are not numbered from the start methionine, but are instead enumerated from downstream sequence. As such, confusion can arise with the non-standard numbering. The CEACAM1 proteins containing the N-terminal IgV-domain has its initial residue 34 amino acids from the start methionine.

Appendix C: Plasmid Maps

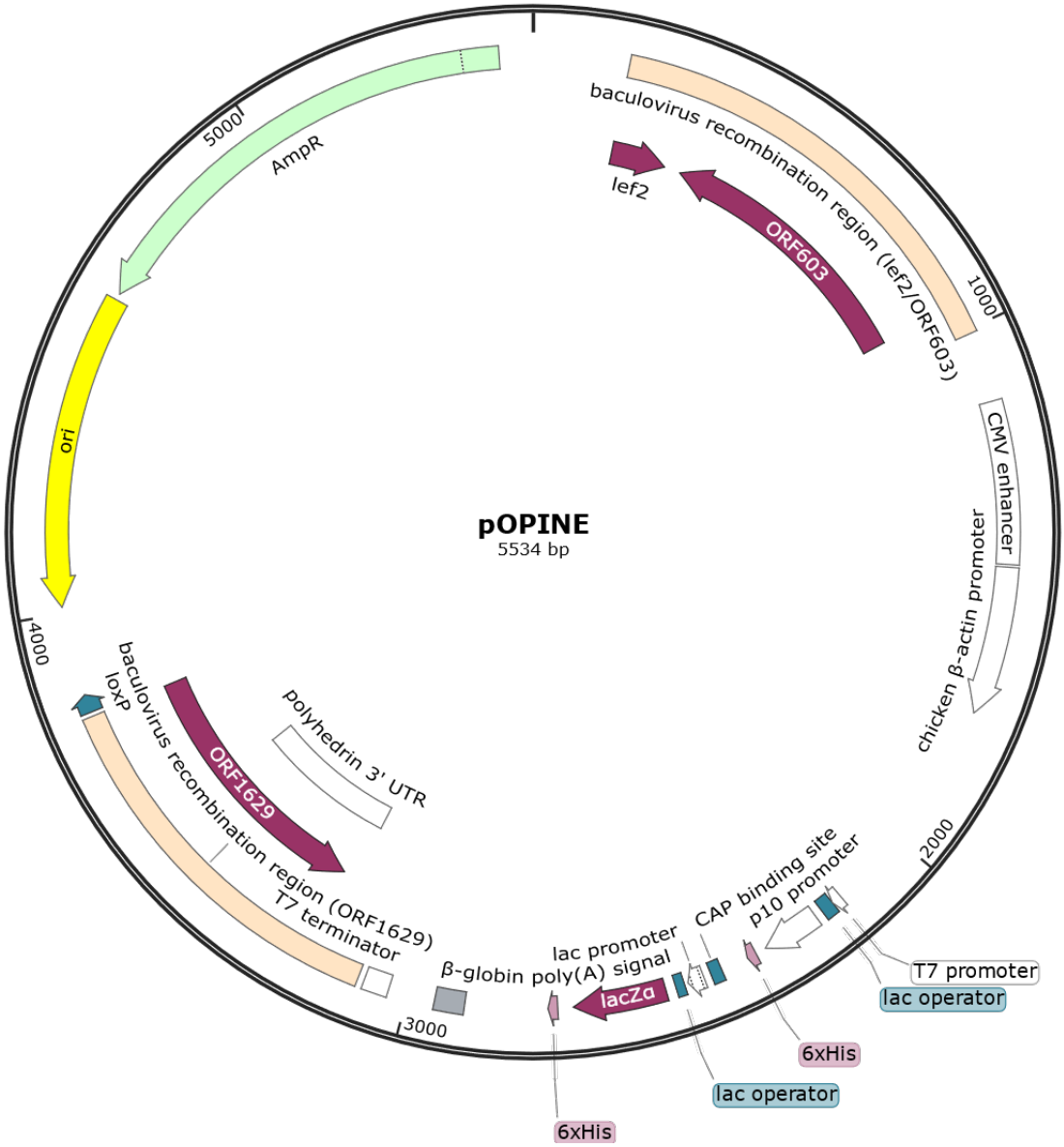


Figure S 1 | **pOPINE plasmid map.**

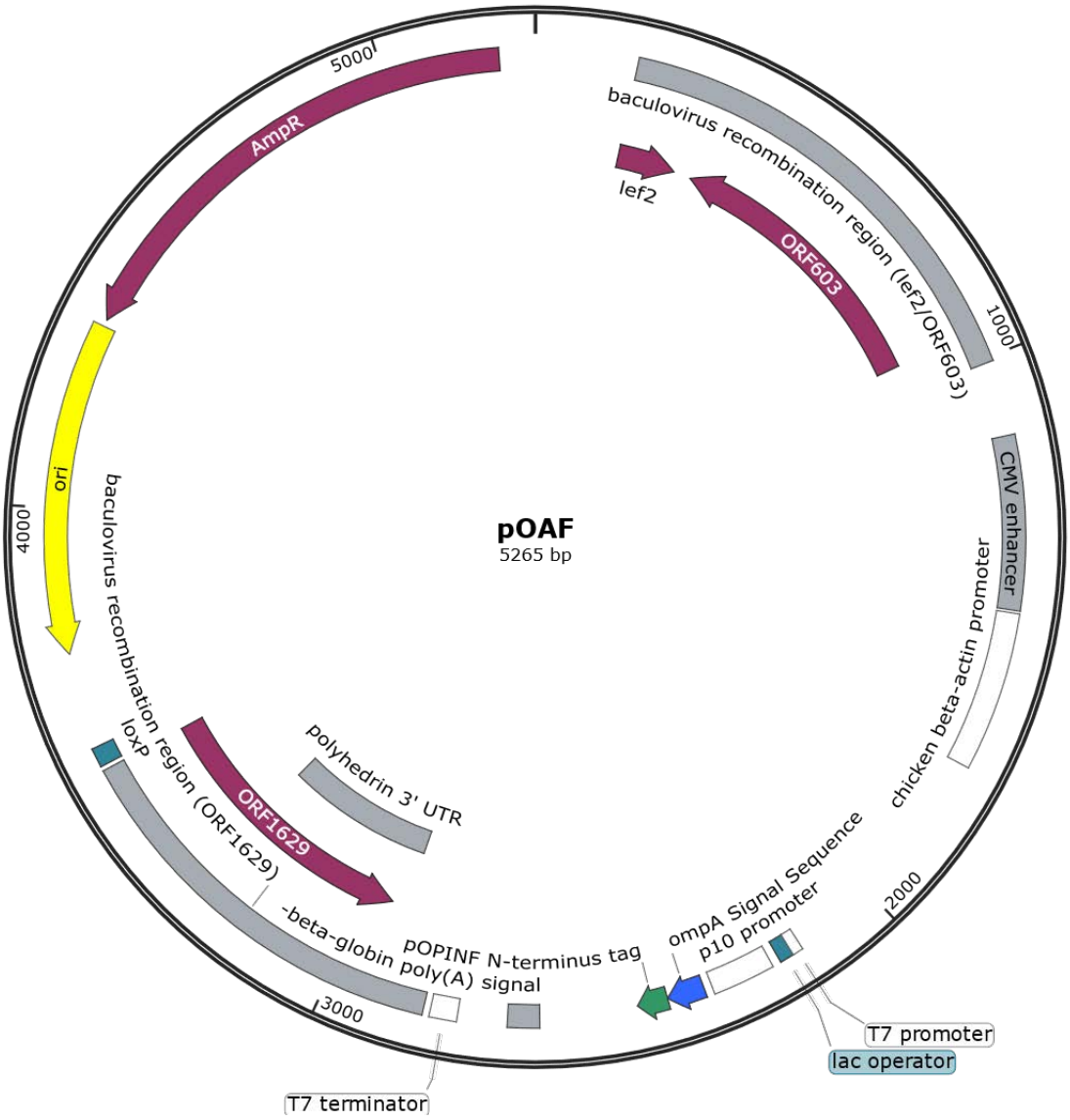


Figure S 2 | **pOAF plasmid map.**

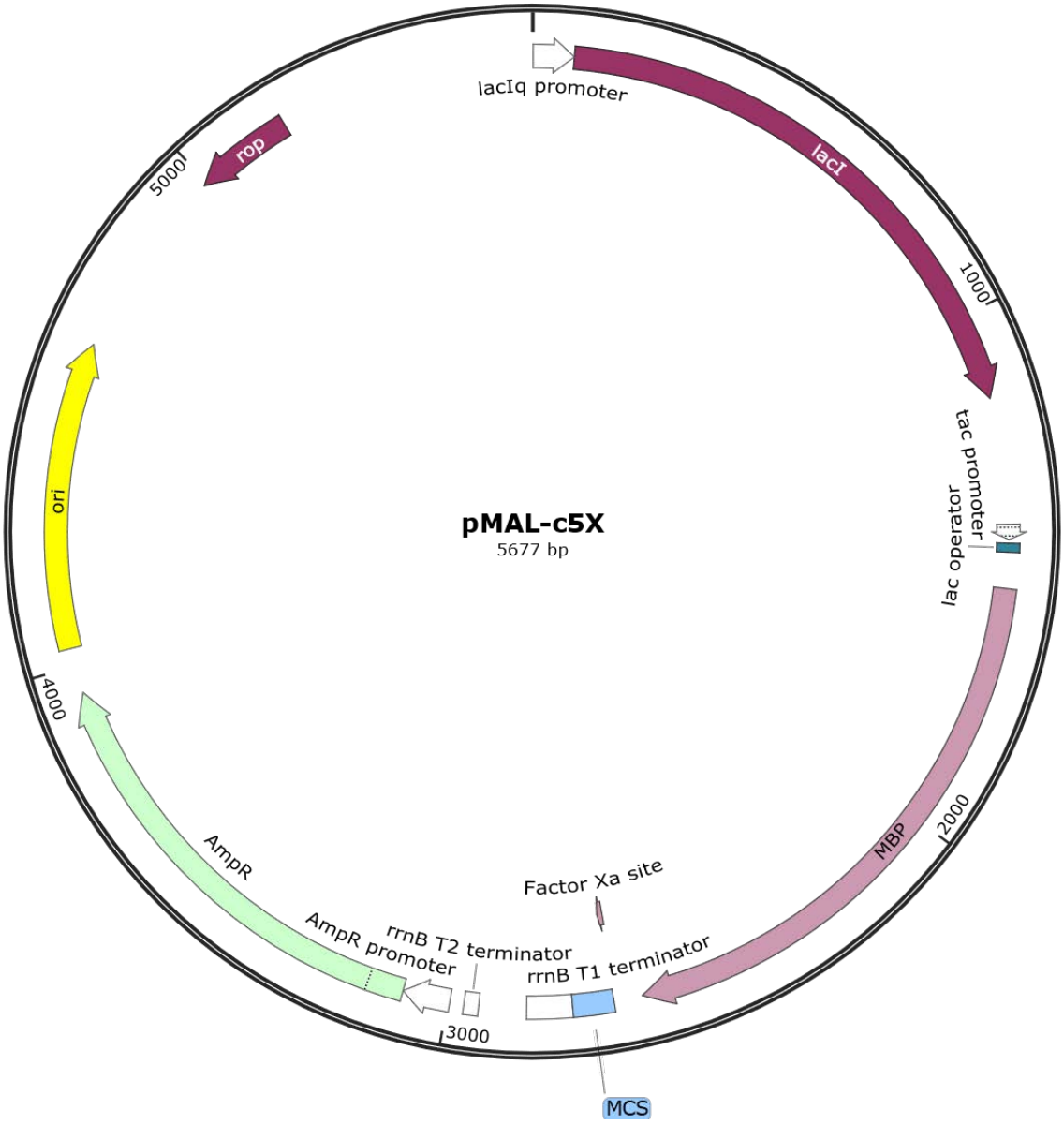


Figure S 3 | **pMAL-c5X plasmid map.**

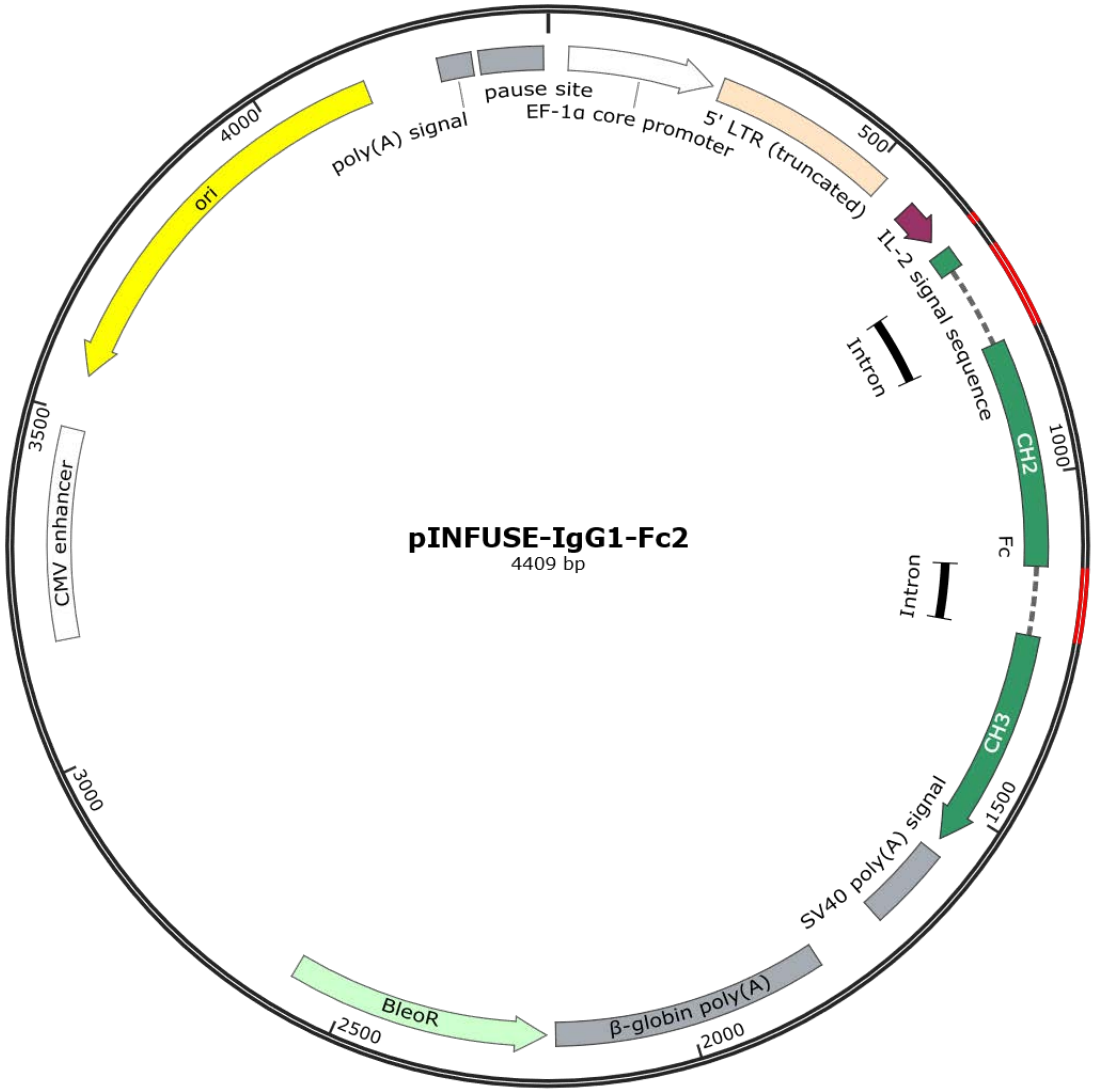


Figure S 4 | **pINFUSE-IgG1-Fc2 plasmid map.**

Appendix D: Gene Sequences and Alignments

>cbpFa [*F. nucleatum* ATCC 25586]
ATGAAAAAATTTGTTAGTTTAAAATTAATTGTTTTTAGTTTTATTTTAGTTGCTGGTAGTGTTTCT
TATTCAGCTGCACCAGTTATTAAAGCAGGAACTGCTACTGATAGTACAGAAGCAGGAGTTGATAAT
GTAGCTAATGGAGTAAAAAGTTCAGCTTTTGGATATGATAATAAAGCCATTGAAAAGGAAAGTTCA
GCTTTTGGAACTGGGAATAGAGCTACTGGTGAGTTTAGTTCAGCTTTTGGATTTCATAATATAGCC
AGTAAAATACATAGCTCAGCTTTTGGAAGCAATAATGCAGCTGATGGGGTAAATAGTTCAGCTTTT
GGATTTAAAAATACAGTTAGTGGATTTAATAGTTCAGCTTTTGGAAGTCAATATCAAGTTACTGGA
AACTTTTCTGGTGCTTTTGGAATGGGTGAATTCAATGGTCAGTATCAATATAAAAATGAAGGTAAT
AATTCATATATGATCGGTAACAAGAATAAAATTGCTAGTGGCTCTGATGATAACTTTATTTTAGGT
AATAATGTTCATATTGGCGGTGGTATTAATAATTCAGTAGCTCTTGGTAATAATTCTACTGTTAGT
GCTTCTAATACTGTTTCTGTTGGATCTTCTACATTAAAAAGAAAGATAGTTAATGTTGGAGATGGA
GCTATTTCTGCTAATTCTAGTGATGCTGTTACTGGTAGACAATTATATAGTGGAAATGGAATTGAT
ACTGCTGCTTGGCAAAATAAATTAAATGTTACTAGAAAAAATGACTATAAAGATGCTAATGATATT
GATGTTAATAAATGGAAGGCAAAACTTGGTGTTGGCTCTGGTGGAGGTGGAGGAGCTCCTGTTGAT
GCTTATACTAAAAGTGAAGCTGATAATAAATTTGCAAATAAAACTGATTTAAATGATTATACTAAA
AAAGATGACTATAAAGATGCTAATGGCATTGATGTTGATAAGTGGAAAGCTAAGCTTGGCACTGGT
GCTGGGACTGCTGATATTGAAAATTTAAGAAATGAAGTAAATGAAAAAATTGATGATGTCAAAGAT
GAAGTTAGAACTGTTGGTTCTTTAAGTGCAGCTCTTGCTGGATTACATCCTATGCAATATGACCCA
AAAGCTCCTGTACAAGTTATGGCTGCATTGGGACATTACAGAGATAAACAATCAGTGGCTGTTGGA
GCAAGTTATTATTTCAATGATAGATTTATGATGAGTACAGGTATTGCTCTTTCAGGAGAAAAGAGA
ACTAAAACTATGGCTAATGTAGGATTTACTTTAAAACTTGGTAAGGGTAGTGGAGTTACTTATGAT
GAAACTCCTCAGTATGTTGTTCAAAATGAAGTTAAGAGATTGACAGTTGAAAATCAAGAATTAAAA
GAAAGAGTTAGAAACTTGGAAGAAAAGTTAAATATGTTATTAAAAAATAAATAG

>cbpFb [*F. oralis* sp. nov. 2B3]
ATGAAAAAATTTGTTAGTTTAAAATTAATTGTTTTTAGTCTACTTTTAGTTACGGGGGGCCTTTCT
TATTCAGCCCCAGCTTTTGGAACAGGAACAGGAGCTAATAGTATAGTAGCAGGAGAAGCTAATGAA
GCCACTCAAGAAAAAAGTTCAGCTATTGGATATGGAAATAAAGCTAATGGAAAGTTTAGTTCAGCT
TTTGGGAATGATAATAAAGCTAGTGGAGAAAATAGTTCAGCCTTTGGACGTTCTAATATAGCCAGT
AATGGAACTAGTTCAGCTTTTGGATATTATAATACAGCCAGTGGACTACGTAGCTCAGCTTTTGGA
CATAATAATACAGCTAGTGGAGAAAATAGTTCAGCCTTTGGATATTTTAACACAGCTAGTGAAGAA
AATACTTCAGCTATTGGATTTAAAAATGAAGCTAGTGGAAAACAAAGTTCAGCTATTGGATATTTG
AATACAGCCAGTGCACTACGTAGCTCAGCTTTTGGAATTAATAATACAGCTAGTGGAGAGGGTAGT
TCAGCTTTTGGATATATTAATAAAGTCAGTGGAGCAAACAGTTCTGTTTTAGGAAACCAATATGAA
GTTACCGGAAACTCTTCAGGTGCTTTTGGGGTAGGTTTTTGGAATTCTGGTAGTCATCTATATAAA
AATGAAGGTAATAATTCATATATGATAGGTAATAAAAATAAAATTGCTAGTGGATCTGATGATAAC
TTTATTTTAGGTAATAATGTTGAGATTGGTGCTGGAGTTCAAAAATCTGTTGTTTTAGGAGATGGT
TCTGCTTCAGGTGGAAGTAATACTGTTTCTGTTGGTTCTTCAACTTTACAAAGAAAAATAGTTAAT
GTTGCAGATGGAACGATCTCTGCTACATCTACTGATGCTGTTACTGGTAGACAATTATACAGTGGG
GATGGCATTGATGTTAATAAGTGGAGAACTAAACTTGGTGTTAGCTCTGGTGGAGGTGCAAGTGGA
GGAGCTCCTGGTGATGCCTATACAAAAAGTGAAGCTGATAATAAATTTACAAGTAAAGATGATTAT
AAAGATGCTAATGGAATTGATGTAGATAAATGGAAAGCTAAACTTGGTACAGGGGGAGGTTCTGCT
GATATTCAAAATTTAAGAAATGAAGTCAATGAAAAAATTGATAATGTTAAAGATGAAGTTAGAGGA
GTAGGATCTTTAAGTGCAGCTCTTGCAGGATTACATCCTATGCAATATGATCCAAAAGCTCCTGCA
CAAGTTATGGCAGCATTAGGACATTATAAAGATAGACAAGCTGTCGCTGTTGGAGCAAGTTATTAT
TTCAATGATAAATTTATGATGAGTACAGGTGTTGCTCTTTCAGGAGAAAAGAGAACTGAAGCAATG

GCTAATGTAGGATTTACTTTAAAAATTGGAAAAGGTAGTGGCACTACTTATACTGAAACTCCTCAA
TATGTTGTTCAAAATGAAGTTAAGAGATTAACAGTTGAAAACCAAGAATTAAAAGAAAGAGTTAGA
AACTTAGAAGAAAAATTAAATATGTTATTAAAAAATAAATAG

>cbpFb [*F. animalis* R5001]
ATGAAAAAATTTGTTAGTTTAAAATTAATTGTTTTTAGTTTTATTTTAGTTGCTGGTAGTGTTTCT
TATTCAGCTACACCAGAAATTAAACAAGGAGATATTGCTGATAGTATAGTAGCAGGAGTTAATAAT
AAAGCCAGTGAACTAGCTAGTTCAGCTTTTGGACATAGTAATACAGCTGAAGGAGCACGTAGTTCA
GCTTTTGGATATAATAATAAAGCTAAGGGAAAAGATAGTTTAGGTTTTGGACATAGTAATACAGCT
GAAGGAGAAAAGAGTTTAGGTTTTGGACATAGTAATACAGCTAAGGGAGCAGAGAGTTTAGCTATT
GGACATTCTAACCTTGCTTTTAAAGAAAAAGCTTCAGCTATTGGATATAAAAATGAAGCTAGTGGA
GAAGTTAGTTCAGCTATTGGATATGTGAATAAAGCTACTGGAGCACGTAGTTCAGCTTTTGGAATT
AATAATACAGCTGATGGAGAAAATAGTTCAGCTTTTGGATTTAAAAATAAAATCAGTGGAAAATGG
AGTTCAGCTTTTGGAAACCAATATGAAGTTACTGGTGAAAAGTCTGGAACATTTGGAGTAGGTGAA
TATAATGGTCAGTATAAATATAAAAATGAAGGTAATAATTCATATATGATAGGAAATTATAATAAA
ATTGCTAAGGACTCTAATGATAACTTTATCTTAGGTAATAATGTTGAGATTGGTGCTGGGGTTCAA
AAATCTGTTGTTTTAGGAGATGGTTCTGCTTCTGGTGGAAGTAATACTGTTTCTGTTGGGTCTTCA
ACTTTACAAAGAAAAATAGTTAATGTTGCAGATGGAACTATTTCTGCTACATCTACTGATGCCGTT
ACTGGTAGACAATTATACAGTGGGGATGGCATCGATGTTAATAAATGGAGAACTAGACTTGGTGTT
GGCTCTGGTGGAGGTGCAGGTGGAGGAGCTCCTGTTGATGCCTATACTAAAAGTGAAGCTGATAAT
AAATTTGCAAATAAAACTGATTTAGATAATTATACTAAAAAAGATGATTATAAAGATGCTAATGGC
ATTGATGTTGATAAATGGAAGGCTAAACTTGGTACAGGAGCTGATTCTGCTGATATTCAAAATTTA
AGAAATGAAGTATATGAAAGAATTGATAATGTTAAAGATGAAGTAAGAGATGTAGGTTCTTTAAGT
GCTGCTCTTGCTGGATTACACCCTATGCAATATGACCCAAAAGCTCCTGCACAAGTTATGGCTGCA
TTAGGACATTATAAAGATAGACAAGCTGTTGCTGTTGGAGCAAGTTATTATTTTAATGATAGATTT
ATGATGAGTACAGGTGTTGCTCTTTCAGGAGAAAAGAAAACTAAAACTATGGCTAATGTAGGCTTT
ACTTTAAAACTTGGTAAGGGTAGTGGAACTACTTATAGTGAAACTCCTCAATATGTTGTTCAAAAT
GAAGTTAAGAGACTAACAGTTGAAAATCAAGAATTAAAAGAAAGACTTAGAAACTTAGAACAAAAA
TTAGAAATCTTATTAAAAAATAAATAA

>cbpFc [*F. ovarium* sp. nov. R16531]
ATGAAAAAAATTGTTAGTTTAAAATTAATTATTTTTAGTTTACTTTTAGTTACTGTTGGTATTTCT
TATTCAGCTCCAGCTATTAATCCAGGAACTGGTACTAATAGTATAATTGCAGGAGAGGACAATAAA
GCTACTAAAGATAAAAGTTCTGCTTTTGGACACAGTAATGAAGCCAATGGAAATGTTAGTTCAGCC
TTTGGATACAAGAATAAAGCCAATGGAGAGCGTAGTTCTGCTTTTGGAACTGCAAATACAGCTGAT
GGAGAAAATAGTTCTGCTTTTGGGATTTTAAATAAAACCAGTGGAAAAAACAGCTCTGTTTTTGGA
AGCCAATATGAAGTCACTGGAGATTCTTCTGGTGCTCTTGGGAAAGGTGAATATAATGGTCAGTAT
CAATATAAAAATGAAGGTCATAATTCATATATGATAGGTAATAAAAATAAAATTGCTAAAGGATCT
AATGATAACTTTATTTTAGGTAATAATGTTTCTATTGGTAAGGGTATTCAAAATTCAGTAGCTCTT
GGTAATAATTCTACTGTTACTGCTTCTAATACTGTTTCTGTTGGTTCTGCTACTTTAAAAAGAAAA
ATAGTTAATGTTGGAGATGGAGAAGTTTCAGCTACTTCATCTGATGCTGTTACTGGTAAACAATTA
TATAGAGGAGAAGGTATTGATGTTAATGCTTGGAGAGCTAAATTAGGTGTTGGTACTGGTTCTGCT
GATATTCAAAATTTAAGAAATGAAGTATATGAAAGAATTGATAATGTAAAAGATGAAGTAAGAGAT
GTAGGTTCTTTAAGTGCTGCTCTTGCTGGATTACACCCTATGCAATATGACCCAAAAGCTCCTGCA
CAAGTTATGGCTGCATTAGGACATTATAAAGATAGACAAGCTGTTGCTGTTGGAGCAAGTTATTAT
TTTAATGATAGATTTATGATGAGTACTGGTGTTGCTCTTTCAGGAGAAAAGAAAACTAAGACTATG
GCTAATGTAGGATTTACTTTAAAACTTGGTAAAGGCAGTGGAACTACTTATAGTGAAACTCCTCAA

TATGTTGTTCAAAATGAAGTTAAGAGACTAACAGTTGAAAATCAAGAATTAAAAGAAAGAGTTAGA
AACTTGGAAGAAAAATTAAATATGTTATTAAAAAGCAAATAA

>CbpFa [*F. nucleatum ATCC 25586*]
MKKFVSLKLIVFSFILVAGSVSYSAAPVIKAGTATDSTEAGVDNVANGVKSSAFGYDNKAIEKESS
AFGTGNRATGEFSSAFGFHNIASKIHSSAFGSNNAADGVNSSAFGFKNTVSGFNSSAFGSQYQVTG
NFSGAFGMGEFNGQYQYKNEGNNSYMIGNKNKIASGSDDNFILGNNVHIGGGINNSVALGNNSTVS
ASNTVSVGSSTLKRKIVNVGDGAISANSSDAVTGRQLYSGNGIDTAAWQNKLNVTRKNDYKDANDI
DVNKWKAKLGVGSGGGGGAPVDAYTKSEADNKFANKTDLNDYTKKDDYKDANGIDVDKWKAKLGTG
AGTADIENLRNEVNEKIDDVKDEVRTVGSLSAALAGLHPMQYDPKAPVQVMAALGHYRDKQSVAVG
ASYYFNDRFMMSTGIALSGEKRTKTMANVGFTLKLGKGSGVTYDETPQYVVQNEVKRLTVENQELK
ERVRNLEEKLNMLLKNK

>CbpFb [*F. oralis* sp. nov. 2B3]
MKKFVSLKLIVFSLLLVTGGLSYSAPAFGTGTGANSIVAGEANEATQEKSSAIGYGNKANGKFSSA
FGNDNKASGENSSAFGRSNIASNGTSSAFGYYNTASGLRSSAFGHNNTASGENSSAFGYFNTASEE
NTSAIGFKNEASGKQSSAIGYLNTASALRSSAFGINNTASGEGSSAFGYINKVSGANSSVLGNQYE
VTGNSSGAFGVGFWNSGSHLYKNEGNNSYMIGNKNKIASGSDDNFILGNNVEIGAGVQKSVVLGDG
SASGGSNTVSVGSSTLQRKIVNVADGTISATSTDAVTGRQLYSGDGIDVNKWRTKLGVSSGGGASG
GAPGDAYTKSEADNKFTSKDDYKDANGIDVDKWKAKLGTGGGSADIQNLRNEVNEKIDNVKDEVRG
VGSLSAALAGLHPMQYDPKAPAQVMAALGHYKDRQAVAVGASYYFNDKFMMSTGVALSGEKRTEAM
ANVGFTLKIGKGSGTTYTETPQYVVQNEVKRLTVENQELKERVRNLEEKLNMLLKNK

>CbpFb [*F. animalis* R5001]
MKKFVSLKLIVFSFILVAGSVSYSATPEIKQGDIADSIVAGVNNKASELASSAFGHSNTAEGARSS
AFGYNNKAKGKDSLGFGHSNTAEGEKSLGFGHSNTAKGAESLAIGHSNLAFKEKASAIGYKNEASG
EVSSAIGYVNKATGARSSAFGINNTADGENSSAFGFKNKISGKWSSAFGNQYEVTGEKSGTFGVGE
YNGQYKYKNEGNNSYMIGNYNKIAKDSNDNFILGNNVEIGAGVQKSVVLGDGSASGGSNTVSVGSS
TLQRKIVNVADGTISATSTDAVTGRQLYSGDGIDVNKWRTRLGVGSGGGAGGGAPVDAYTKSEADN
KFANKTDLDNYTKKDDYKDANGIDVDKWKAKLGTGADSADIQNLRNEVYERIDNVKDEVRDVGSLS
AALAGLHPMQYDPKAPAQVMAALGHYKDRQAVAVGASYYFNDRFMMSTGVALSGEKKTKTMANVGF
TLKLGKGSGTTYSETPQYVVQNEVKRLTVENQELKERLRNLEQKLEILLKNK

>CbpFc [*F. ovarium* sp. nov. R16531]
MKKIVSLKLIIFSLLLVTVGISYSAPAINPGTGTNSIIAGEDNKATKDKSSAFGHSNEANGNVSSA
FGYKNKANGERSSAFGTANTADGENSSAFGILNKTSGKNSSVFGSQYEVTGDSSGALGKGEYNGQY
QYKNEGHNSYMIGNKNKIAKGSNDNFILGNNVSIGKGIQNSVALGNNSTVTASNTVSVGSATLKRK
IVNVGDGEVSATSSDAVTGKQLYRGEGIDVNAWRAKLGVGTGSADIQNLRNEVYERIDNVKDEVRD
VGSLSAALAGLHPMQYDPKAPAQVMAALGHYKDRQAVAVGASYYFNDRFMMSTGVALSGEKKTKTM
ANVGFTLKLGKGSGTTYSETPQYVVQNEVKRLTVENQELKERVRNLEEKLNMLLKSK

>CEACAM1 [*H. sapiens*]
MGHLSAPLHRVRVPWQGLLLTASLLTFWNPPTTAQLTTESMPFNVAEGKEVLLLVHNLPQQLFGYS
WYKGERVDGNRQIVGYAIGTQQATPGPANSGRETIYPNASLLIQNVTQNDTGFYTLQVIKSDLVNE
EATGQFHVYPELPKPSISSNNSNPVEDKDAVAFTCEPETQDTTYLWWINNQSLPVSPRLQLSNGNR
TLTLLSVTRNDTGPYECEIQNPVSANRSDPVTLNVTYGPDTPTISPSDTYYRPGANLSLSCYAASN
PPAQYSWLINGTFQQSTQELFIPNITVNNSGSYTCHANNSVTGCNRTTVKTIIVTELSPVVAKPQI

KASKTTVTGDKDSVNLTCSTNDTGISIRWFFKNQSLPSSERMKLSQGNTTLSINPVKREDAGTYWC
EVFNPISKNQSDPIMLNVNYNALPQENGLSPGAIAGIVIGVVALVALIAVALACFLHFGKTGRASD
QRDLTEHKPSVSNHTQDHSNDPPNKMNEVTYSTLNFEAQQPTQPTSASPSLTATEIIYSEVKKQ

>CEACAM3 [*H. sapiens*]
MGPPSASPHRECIPWQGLLLTASLLNFWNPPTTAKLTIESMPLSVAEGKEVLLLVHNLPQHLFGYS
WYKGERVDGNSLIVGYVIGTQQATPGAAYSGRETIYTNASLLIQNVTQNDIGFYTLQVIKSDLVNE
EATGQFHVYQENAPGLPVGAVAGIVTGVLVGVALVAALVCFLLLAKTGRTSIQRDLKEQQPQALAP
GRGPSHSSAFSMSPLSTAQAPLPNPRTAASIYEELLKHDTNIYCRMDHKAEVAS

>CEACAM5 [*H. sapiens*]
MESPSAPPHRWCIPWQRLLLTASLLTFWNPPTTAKLTIESTPFNVAEGKEVLLLVHNLPQHLFGYS
WYKGERVDGNRQIIGYVIGTQQATPGPAYSGREIIYPNASLLIQNIIQNDTGFYTLHVIKSDLVNE
EATGQFRVYPELPKPSISSNNSKPVEDKDAVAFTCEPETQDATYLWWVNNQSLPVSPRLQLSNGNR
TLTLFNVTRNDTASYKCETQNPVSARRSDSVILNVLYGPDAPTISPLNTSYRSGENLNLSCHAASN
PPAQYSWFVNGTFQQSTQELFIPNITVNNSGSYTCQAHNSDTGLNRTTVTTITVYAEPPKPFITSN
NSNPVEDEDAVALTCEPEIQNTTYLWWVNNQSLPVSPRLQLSNDNRTLTLLSVTRNDVGPYECGIQ
NELSVDHSDPVILNVLYGPDDPTISPSYTYYRPGVNLSLSCHAASNPPAQYSWLIDGNIQQHTQEL
FISNITEKNSGLYTCQANNSASGHSRTTVKTITVSAELPKPSISSNNSKPVEDKDAVAFTCEPEAQ
NTTYLWWVNGQSLPVSPRLQLSNGNRTLTLFNVTRNDARAYVCGIQNSVSANRSDPVTLDVLYGPD
TPIISPPDSSYLSGANLNLSCHSASNPSPQYSWRINGIPQQHTQVLFIAKITPNNNGTYACFVSNL
ATGRNNSIVKSITVSASGTSPGLSAGATVGIMIGVLVGVALI

>CEACAM8 [*H. sapiens*]
MGPISAPSCRWRIPWQGLLLTASLFTFWNPPTTAQLTIEAVPSNAAEGKEVLLLVHNLPQDPRGYN
WYKGETVDANRRIIGYVISNQQITPGPAYSNRETIYPNASLLMRNVTRNDTGSYTLQVIKLNLMSE
EVTGQFSVHPETPKPSISSNNSNPVEDKDAVAFTCEPETQNTTYLWWVNGQSLPVSPRLQLSNGNR
TLTLLSVTRNDVGPYECEIQNPASANFSDPVTLNVLYGPDAPTISPSDTYYHAGVNLNLSCHAASN
PPSQYSWSVNGTFQQYTQKLFIPNITTKNSGSYACHTTNSATGRNRTTVRMITVSDALVQGSSPGL
SARATVSIMIGVLARVALI
>CEACAM1b [*M. musculus*]
MELASAHLHKGQVPWVGLLLTASLLASWSPPTTAEVTIEAVPPQVAEDNNVLLLVHNLPLALGAFA
WYKGNPVSTNAEIVHFVTGTNKTTTGPAHSGRETVYSNGSLLIQRVTVKDTGVYTIEMTDENFRRT
EATVQFHVHQPVTQPSLQVTNTTVKELDSVTLTCLSNDIGANIQWLFNSQSLQLTERMTLSQNNSI
LRIDPIKREDAGEYQCEISNPVSVKRSNSIKLDIIFDPTQGGLSDGAIAGIVIGVVAGVALIAGLA
YFLYSRKSGGGSDQRDLTEHKPSTSNHNLAPSDNSPNKVDDVAYTVLNFNSQQPNRPTSAPSSPRA
TETVYSEVKKK

All CbpFs Alignment:

```
                    1         11        21        31        41
CbpFa
>GCA_000007325.1    MKKFVSLKLIVFSFILVAGSVSYSAAPVIKAG-TATDSTEAGVDNVANGV   49
>GCA_000163915.2    MKKFVSLKLIVFSFILVAGSVSFS-ADAEFKKENGTDSIIAGISNEASGN   49
>R18528             MKKFVSLKLIVFSFILVAGSVSYSAAPVIKAG-TATDSTEAGVDNVANGV   49
>R28211             MKKFVSLKLIVFSFILVVGSVSFS-ADAEFKKENGTDSIIAGISNEASGN   49
>R32935             MKKFVSLKLIVFSFILVAGSVSYSAAPVIKAG-TATDSTEAGVDNVANGV   49
>GCA_000455945.1    MKKFVSLKLIVFSFILVAGSVSFS-VPVFQAG-TGTDSTVAGVNNEANGE   48
>GCA_000178895.1    MKKFVSLKLIVFSFILVAGSVSFS-APVFQAG-TGTDSTVAGVNNEANGE   48
>GCA_001510735.1    MKKFVSLKLIVFSFILVAGSVSFS-APVFQAG-TGTDSTVAGVNNEANGE   48
>GCA_001296165.1    MKKFVSLKLIVFSFILVAGSVSFS-VPVFQAG-TGTDSTVAGVNNEANGE   48
>GCA_001296185.1    MKKFVSLKLIVFSFILVAGSVSFS-APVFQAG-TGTDSTVAGVNNEANGE   48
>GCA_002211605.1    MKKFVSLKLIVFSFILVAGSVSFS-VPVFQAG-TGTDSTVAGVNNEANGE   48
>R24394             MKKFVSLKLIVFSFILVAGSVSFS-APVFQAG-TGTDSTVAGVNNEANGE   48
>GCA_000158255.2    MKKFVSLKLIVFSFILVASSVSFS-VPAIQGG-TGSDSTVAGIGNDAS--   46
>GCA_000162235.2    MKKFVSLKLIVFSFILVAGSVSFS-APGIHAG-TVTDSIIAGIDNIAS--   46
>GCA_000182945.1    MKKFVSLKLIVFSFILVASSVSFS-VPVIQGG-TGSDSTVAGIGNDAS--   46
>GCA_000347315.1    MKKFVSLKLIVFSFILVAGSVSFS-APGIHAG-TVTDSIIAGIDNIAS--   46
>GCA_000479205.1    MKKFVSLKLIVFSFILVTGSVSFS-VPAIQGG-TGTDSIIAGIDNIAS--   46
>GCA_000517705.1    MKKFVSLKLIVFSFILVTGSVSFS-VPAIQGG-TGTDSIIAGIDNIAS--   46
>GCA_001296125.1    MKKFVSLKLIVFSFILVTGSVSFS-VPAIQGG-TGTDSIIAGIDNIAS--   46
>GCA_001810995.1    MKKFVSLKLIVFSFILVASSVSFS-VPAIQGG-TGSDSTVAGIGNDAS--   46
>GCA_001854465.1    MKKFVSLKLIVFSFILVTGSVSFS-VPAIQGG-TGTDSIIAGIDNIAS--   46
>GCA_002749995.1    MKKFVSLKLIVFSFILVTGSVSFS-VPAIQGG-TGTDSIIAGIDNIAS--   46
>GCA_002764055.1    MKKFVSLKLIVFSFILVASSVSFS-VPAIQGG-TGSDSTVAGIGNDAS--   46
>GCA_000455965.1    MKKFVSLKLIVFSFILVTGSVSFS-VPAIQGG-TGTDSIIAGIDNIAS--   46
>R26872             MKKFVSLKLIVFSFILVAGSVSFS-APGIHAG-TGTDSIIAGIDNIAS--   46
>R28385             MKKFVSLKLIVFSFILVASSVSFS-VPAIQGG-TGSDSTIAGVDNIAS--   46
>R28400             MKKFVSLKLIVFSFILVASSVSFS-VPAIQGG-TGSDSTIAGIENDAS--   46
>2B17               MKKFVSLKLIVFSFILVASSVSFS-VPAIQGG-TGSDSTVAGIGNDAS--   46
>R29976             MKKFVSLKLIVFSFILVTGSVSFS-VPAIQGG-TGSDSTVAGIGNDAS--   46
>R30464             MKKFVSLKLIVFSFILVTGSVSFS-VPAIQGG-TGTDSIIAGIDNIAS--   46
>R30604             MKKFVSLKLIVFSFILVAGSVSFS-APGIHAG-TVTDSIIAGIDNIAS--   46
>R31249             MKKFVSLKLIVFSFILVASSVSFS-VPAIQGG-TGSDSTVAGIGNDAS--   46
>R32310             MKKFVSLKLIVFSFILVASSVSFS-VPAIQGG-TGSDSTVAGIGNDAS--   46
>R33458             MKKFVSLKLIVFSFILVASSVSFS-VPAIQGG-TGSDSTVAGIGNDAS--   46
>R33533             MKKFVSLKLIVFSFILVTGSVSFS-APGIQGG-TGSDSTVAGIGNDAS--   46
>GCA_002243405.1    MKKFVSLKLIVFSFILVVGSVSFS-APVFQGG-TGTDSTVAGVDNIAS--   46
CbpFb
>2B16               MKKFVSLKLIVFSLLLVTGGLSYSA-PAIGSG-TGANSIVAGETNEAT--   46
>2B2                MKKFVSLKLIVFSLLLVTGGLSYSA-PAIGSG-TGANSIVAGLNNTAD--   46
>2B3                MKKFVSLKLIVFSLLLVTGGLSYSA-PAFGTG-TGANSIVAGEANEAT--   46
>2B4                MKKFVSLKLIVFSLLLVTGGLSYSA-PAIGSG-TGANSIVAGLNNTAD--   46
>R5001              MKKFVSLKLIVFSFILVAGSVSYSATPEIKQG-DIADSIVAGVNNKAS--   47
>GCA_001546435.1    MKKFISLKLIVFSFILVTSAVSYSN-PKIEEG-TVADSIKAGLKNAAD--   46
>R15792             MKKFISLKLIVFSFILVTSAVSYSN-PKIEAG-TGANSIKAGLDNEAD--   46
CbpFc
>R16531_1           MKKIVSLKLIIFSLLLVTVGISYS-APAINPG-TGTNSIIAGEDNKAT--   46
```

```
                    51        61        71        81        91
CbpFa
>GCA_000007325.1    KSSAFGYDNKAIEKES----------------------SAFGTG----   71
>GCA_000163915.2    ESSAFGYKNKATEEKS----------------------SAFGHS----   71
>R18528             KSSAFGYDNKAIEKES----------------------SAFGTG----   71
>R28211             ESSAFGYKNKATEEKS----------------------SAFGHS----   71
>R32935             KSSAFGYDNKAIEKES----------------------SAFGTG----   71
>GCA_000455945.1    KSSAFGYENKAKEKLSSAFGYKNIANGIEG----------SAFGIS----   84
>GCA_000178895.1    KSSAFGYENKAKEKLSSAFGYKNIANGIEG----------SAFGIS----   84
>GCA_001510735.1    KSSAFGYENKAKEKLSSAFGYKNIANGIEG----------SAFGIS----   84
>GCA_001296165.1    KSSAFGYENKAKEKLSSAFGYKNIANGIEG----------SAFGIS----   84
>GCA_001296185.1    KSSAFGYENKAKEKLSSAFGYKNIANGIEG----------SAFGIS----   84
>GCA_002211605.1    KSSAFGYENKAKEKLSSAFGYKNIANGIEG----------SAFGIS----   84
>R24394             KSSAFGYENKAKEKLSSAFGYKNIANGIEG----------SAFGIS----   84
>GCA_000158255.2    ------------GENS----------------------SAFGFH----   56
>GCA_000162235.2    ------------EEKS----------------------SAFGFK----   56
>GCA_000182945.1    ------------GENS----------------------SAFGFH----   56
>GCA_000347315.1    ------------EEKS----------------------SAFGFK----   56
>GCA_000479205.1    ------------EEKS----------------------SAFGFK----   56
>GCA_000517705.1    ------------EEKS----------------------SAFGFK----   56
>GCA_001296125.1    ------------EEKS----------------------SAFGFK----   56
>GCA_001810995.1    ------------GENS----------------------SAFGFH----   56
>GCA_001854465.1    ------------EEKS----------------------SAFGFK----   56
>GCA_002749995.1    ------------EEKS----------------------SAFGFK----   56
>GCA_002764055.1    ------------GENS----------------------SAFGFH----   56
>GCA_000455965.1    ------------EEKS----------------------SAFGFK----   56
>R26872             ------------EEKS----------------------SAFGFK----   56
>R28385             ------------EEKS----------------------SAFGFK----   56
>R28400             ------------GENS----------------------SAFGFH----   56
>2B17               ------------GENS----------------------SAFGFH----   56
>R29976             ------------GENS----------------------SAFGFH----   56
>R30464             ------------EEKS----------------------SAFGFK----   56
>R30604             ------------EEKS----------------------SAFGFK----   56
>R31249             ------------GENS----------------------SAFGFH----   56
>R32310             ------------GENS----------------------SAFGFH----   56
>R33458             ------------GENS----------------------SAFGFH----   56
>R33533             ------------GENS----------------------SAFGFH----   56
>GCA_002243405.1    ------------EEKS----------------------SAFGTG----   56
CbpFb
>2B16               ------------QEKSSAIG----------YGNKAKGKFSSAIGYDNIAS   74
>2B2                ------------KEKSSAFG----------YGNKANGKYSSSFGYDNTAS   74
>2B3                ------------QEKSSAIG----------YGNKANGKFSSAFGNDNKAS   74
>2B4                ------------KEKSSAFG----------YGNKANGKYSSSFGYDNTAS   74
>R5001              ------------ELASSAFG----------HSNTAEGARSSAFGYNNKAK   75
>GCA_001546435.1    ------------KEQSLAFG----------YFNRALGKKSSAFGNA----   70
>R15792             ------------GEESSAFG----------FHNITDKSGSSAFGNG----   70
CbpFc
>R16531_1           ------------KDKS----------------------SAFGHS----   56
```

```
                    101       111       121       131       141
CbpFa
>GCA_000007325.1    ------------------------------------------------   71
>GCA_000163915.2    ------------------------------------------------   71
>R18528             ------------------------------------------------   71
>R28211             ------------------------------------------------   71
>R32935             ------------------------------------------------   71
>GCA_000455945.1    ------------------------------------------------   84
>GCA_000178895.1    ------------------------------------------------   84
>GCA_001510735.1    ------------------------------------------------   84
>GCA_001296165.1    ------------------------------------------------   84
>GCA_001296185.1    ------------------------------------------------   84
>GCA_002211605.1    ------------------------------------------------   84
>R24394             ------------------------------------------------   84
>GCA_000158255.2    ------------------------------------------------   56
>GCA_000162235.2    ------------------------------------------------   56
>GCA_000182945.1    ------------------------------------------------   56
>GCA_000347315.1    ------------------------------------------------   56
>GCA_000479205.1    ------------------------------------------------   56
>GCA_000517705.1    ------------------------------------------------   56
>GCA_001296125.1    ------------------------------------------------   56
>GCA_001810995.1    ------------------------------------------------   56
>GCA_001854465.1    ------------------------------------------------   56
>GCA_002749995.1    ------------------------------------------------   56
>GCA_002764055.1    ------------------------------------------------   56
>GCA_000455965.1    ------------------------------------------------   56
>R26872             ------------------------------------------------   56
>R28385             ------------------------------------------------   56
>R28400             ------------------------------------------------   56
>2B17               ------------------------------------------------   56
>R29976             ------------------------------------------------   56
>R30464             ------------------------------------------------   56
>R30604             ------------------------------------------------   56
>R31249             ------------------------------------------------   56
>R32310             ------------------------------------------------   56
>R33458             ------------------------------------------------   56
>R33533             ------------------------------------------------   56
>GCA_002243405.1    ------------------------------------------------   56
CbpFb
>2B16               GLDSSAFGRSNIADKEGSSAIGYYNTASGKNSSSFGYKNTASGENSSAFG   124
>2B2                GLDSSAFGRSNIADKEASSAIGYYNTASGKNSSSFGYKNTASGENSSAFG   124
>2B3                GENSSAFGRSNIASNGTSSAFGYYNTASGLRSSAFGHNNTASGENSSAFG   124
>2B4                GLDSSAFGRSNIADKEASSAIGYYNTASGKNSSSFGYKNTASGENSSAFG   124
>R5001              GKDSLGFGHS----------NTAEGEKSLGFGHSNTAKGAESLAIG      111
>GCA_001546435.1    ------------------------------------------------   70
>R15792             ------------------------------------------------   70
CbpFc
>R16531_1           ------------------------------------------------   56
```

```
                    151       161       171       181       191
CbpFa
>GCA_000007325.1    ----------------NRATGEFSSAFGFHNIASKIHSSAFGSNNAADGV   105
>GCA_000163915.2    ----------------NEASGKFSSAFGYMNEANGKYSSAFGTKNIASEE   105
>R18528             ----------------NRATGEFSSAFGFHNIASKIHSSAFGSNNAADGV   105
>R28211             ----------------NEASGKFSSAFGYKNIASSLRSSAFGVGNKASGN   105
>R32935             ----------------NRATGEFSSAFGFHNIASKIHSSAFGSNNAADGV   105
>GCA_000455945.1    ----------------NLAKGQYSSAFGFRNVANKRHSSAFGSGNEANGE   118
>GCA_000178895.1    ----------------NLAKGQYSSAFGFRNVANKRHSSAFGSGNEANGE   118
>GCA_001510735.1    ----------------NLAKGQYSSAFGFRNVANKRHSSAFGSGNEANGE   118
>GCA_001296165.1    ----------------NLAKGQYSSAFGFRNVANKRHSSAFGSGNEANGE   118
>GCA_001296185.1    ----------------NLAKGQYSSAFGFRNVANKRHSSAFGSGNEANGE   118
>GCA_002211605.1    ----------------NLAKGQYSSAFGFRNVANKRHSSAFGSGNEANGE   118
>R24394             ----------------NLAKGQYSSAFGFRNVANKRHSSAFGSGNEANGE   118
>GCA_000158255.2    ----------------NTASGKFSSAFGFRNIASKLRSSAFGTGNKADGE   90
>GCA_000162235.2    ----------------NKASGKFSSAFGYMNEANGKYSSAFGTKNIASEE   90
>GCA_000182945.1    ----------------NTASGKFSSAFGFRNIASKLRSSAFGTGNKADGE   90
>GCA_000347315.1    ----------------NKASGKFSSAFGYMNEANGKYSSAFGTKNIASEE   90
>GCA_000479205.1    ----------------NKASGKFSSAFGYMNEANGQYSSAFGAGNKASGE   90
>GCA_000517705.1    ----------------NKASGKFSSAFGYMNEANGQYSSAFGAGNKASGE   90
>GCA_001296125.1    ----------------NKASGKFSSAFGFRNIASKLRSSAFGTGNKADGE   90
>GCA_001810995.1    ----------------NTASGKFSSAFGFRNIASKLRSSAFGTGNKADGE   90
>GCA_001854465.1    ----------------NKASGKFSSAFGYMNEANGQYSSAFGAGNKASGE   90
>GCA_002749995.1    ----------------NKASGKFSSAFGFRNIASKLRSSAFGTGNKADGE   90
>GCA_002764055.1    ----------------NTASGKFSSAFGFRNIASKLRSSAFGTGNKADGE   90
>GCA_000455965.1    ----------------NKASGKFSSAFGFRNIASKLRSSAFGTGNKADGE   90
>R26872             ----------------NKASGKFSSAFGYMNEANGKYSSAFGTKNIASEE   90
>R28385             ----------------NKASGKFSSAFGYMNEANGKYSSAFGTKNIASEE   90
>R28400             ----------------NKASGKFSSAFGFRNIASKLRSSAFGTGNKADGE   90
>2B17               ----------------NTASGKFSSAFGFRNIASKLRSSAFGTGNKADGE   90
>R29976             ----------------NTASGKFSSAFGFRNIASKLRSSAFGTGNKADGE   90
>R30464             ----------------NKASGKFSSAFGYMNEANGQYSSAFGAGNKASGE   90
>R30604             ----------------NKASGKFSSAFGYMNEANGKYSSAFGTKNIASEE   90
>R31249             ----------------NTASGKFSSAFGFRNIASKLRSSAFGTGNKADGE   90
>R32310             ----------------NTASGKFSSAFGFRNIASKLRSSAFGTGNKADGE   90
>R33458             ----------------NTASGKFSSAFGFRNIASKLRSSAFGTGNKADGE   90
>R33533             ----------------NTASGKFSSAFGFRNIASKLRSSAFGTGNKADGE   90
>GCA_002243405.1    ----------------NNASGKFSSAFGYKNIASRLRSSAFGTGNKAIGE   90
CbpFb
>2B16               YFNTASEENTSAIGFKNEASGKQSSAFGYLNKASALRSSAFGINNTADGE   174
>2B2                YFNTASEENTSAIGFKNEASGKQSSAIGYLNTASALRSSAFGINNTADGE   174
>2B3                YFNTASEENTSAIGFKNEASGKQSSAIGYLNTASALRSSAFGINNTASGE   174
>2B4                YFNTASEENTSAIGFKNEASGKQSSAIGYLNTASALRSSAFGINNTADGE   174
>R5001              HSNLAFKEKASAIGYKNEASGEVSSAIGYVNKATGARSSAFGINNTADGE   161
>GCA_001546435.1    ----------------NIAVGENSSAFGYHNIANNNQSSAFGFGNKSIGE   104
>R15792             ----------------NVAIGENSSAFGFHNIASKIHSSAFGSSNEVDGD   104
CbpFc
>R16531_1           ----------------NEANGNVSSAFGYKNKANGERSSAFGTANTADGE   90
```

```
                    201       211       221       231       241
CbpFa
>GCA_000007325.1    NSSAFGFKNTVS-----------GFNSSAFGSQYQVTGNFSGAFGMGEF   143
>GCA_000163915.2    QSSAFGFLNKAS-----------GGKSSVFGSQYEVTGNSSGAFGVGEY   143
>R18528             NSSAFGFKNTVS-----------GFNSSAFGSQYQVTGNFSGAFGMGEF   143
>R28211             ESSAFGFLNKAS-----------GGKSSVFGSQYEVTGNSSGAFGVGEY   143
>R32935             NSSAFGFKNTVS-----------GFNSSAFGSQYQVTGNFSGAFGMGEF   143
>GCA_000455945.1    QSSAFGFKNTVS-----------GFNSSAFGSQYQVTGNFSGAFGMGEF   156
>GCA_000178895.1    QSSAFGFKNTVS-----------GFNSSAFGSQYEVTGNFSGAFGMGEF   156
>GCA_001510735.1    QSSAFGFKNTVS-----------GFNSSAFGSQYEVTGNFSGAFGMGEF   156
>GCA_001296165.1    QSSAFGFKNTVS-----------GFNSSAFGSQYQVTGNFSGAFGMGEF   156
>GCA_001296185.1    QSSAFGFKNTVS-----------GFNSSAFGSQYEVTGNFSGAFGMGEF   156
>GCA_002211605.1    QSSAFGFKNTVS-----------GFNSSAFGSQYQVTGNFSGAFGMGEF   156
>R24394             QSSAFGFKNTVS-----------GFNSSAFGSQYEVTGNFSGAFGMGEF   156
>GCA_000158255.2    DSSAFGSLNIAS-----------GKFSSAFGSQYEVTGNSSGAFGVGEF   128
>GCA_000162235.2    QSSAFGFLNKAS-----------GGKSSVFGSQYEVTGNSSGAFGVGEY   128
>GCA_000182945.1    DSSAFGSLNIAS-----------GKFSSAFGSQYEVTGNSSGAFGVGEF   128
>GCA_000347315.1    QSSAFGFLNKAS-----------GGKSSVFGSQYEVTGNSSGAFGVGEY   128
>GCA_000479205.1    QSSAFGFLNKAS-----------GRKSSVFGSQYEVTGNSSGAFGVGEF   128
>GCA_000517705.1    QSSAFGFLNKAS-----------GRKSSVFGSQYEVTGNSSGAFGVGEF   128
>GCA_001296125.1    DSSAFGSLNIAS-----------GKFSSAFGSQYEVTGNSSGAFGVGEF   128
>GCA_001810995.1    DSSAFGSLNIAS-----------GKFSSAFGSQYEVTGNSSGAFGVGEF   128
>GCA_001854465.1    QSSAFGFLNKAS-----------GRKSSVFGSQYEVTGNSSGAFGVGEF   128
>GCA_002749995.1    DSSAFGSLNIAS-----------GKFSSAFGSQYEVTGNSSGAFGVGEF   128
>GCA_002764055.1    DSSAFGSLNIAS-----------GKFSSAFGSQYEVTGNSSGAFGVGEF   128
>GCA_000455965.1    DSSAFGSLNIAS-----------GKFSSAFGSQYEVTGNSSGAFGVGEF   128
>R26872             QSSAFGFLNKAS-----------GGKSSVFGSQYEVTGNSSGAFGVGEY   128
>R28385             QSSAFGFLNKAS-----------GGKSSVFGSQYEVTGNSSGAFGVGEY   128
>R28400             DSSAFGSLNIAS-----------GKFSSAFGSQYEVTGNSSGAFGVGEY   128
>2B17               DSSAFGSLNIAS-----------GKFSSAFGSQYEVTGNSSGAFGVGEF   128
>R29976             DSSAFGSLNIAS-----------GKFSSAFGSQYEVTGNSSGAFGVGEF   128
>R30464             QSSAFGFLNKAS-----------GRKSSVFGSQYEVTGNSSGAFGVGEF   128
>R30604             QSSAFGFLNKAS-----------GGKSSVFGSQYEVTGNSSGAFGVGEY   128
>R31249             DSSAFGSLNIAS-----------GKFSSAFGSQYEVTGNSSGAFGVGEF   128
>R32310             DSSAFGSLNIAS-----------GKFSSAFGSQYEVTGNSSGAFGVGEF   128
>R33458             DSSAFGSLNIAS-----------GKFSSAFGSQYEVTGNSSGAFGVGEF   128
>R33533             DSSAFGSLNIAS-----------GKFSSAFGSQYEVTGNSSGAFGVGEF   128
>GCA_002243405.1    DSSAFGSLNIAS-----------GKFSSAFGSQYEVTGNSSGAFGVGEF   128
CbpFb
>2B16               NSSAFGFKNIVS-----------GFNSSAFGSQYEVTGNFSGAFGMGEF   212
>2B2                NSSAFGFKNKIS-----------GKWSSAFGNQYEVTGEKSGTFGVGEY   212
>2B3                GSSAFGYINKVS-----------GANSSVLGNQYEVTGNSSGAFGVGFW   212
>2B4                NSSAFGFKNKIS-----------GKWSSAFGNQYEVTGEKSGTFGVGEY   212
>R5001              NSSAFGFKNKIS-----------GKWSSAFGNQYEVTGEKSGTFGVGEY   199
>GCA_001546435.1    QSSAFGSLNVVGKLKSDGNPDENYGKKSLAFGSEYEVTGNSSGAFGVGHW   154
>R15792             FSSAFGVKNKIS-----------GKWSSAFGNQYEVTGEKSGTFGVGEY   142
CbpFc
>R16531_1           NSSAFGILNKTS-----------GKNSSVFGSQYEVTGDSSGALGKGEY   128
```

```
                    251       261       271       281       291
CbpFa
>GCA_000007325.1    NG-QYQYKNEGNNSYMIGNKNKIASGSDDNFILGNNVHIGGGINNSVALG   192
>GCA_000163915.2    NG-SYQYKNEGNNSYMIGNKNKIASGSDDNFILGNNVHIGGGINNSVALG   192
>R18528             NG-QYQYKNEGNNSYMIGNKNKIASGSDDNFILGNNVHIGGGINNSVALG   192
>R28211             NG-SYQYKNEGNNSYMIGNKNKIASGSDDNFILGNNVHIGGGINNSVALG   192
>R32935             NG-QYQYKNEGNNSYMIGNKNKIASGSDDNFILGNNVHIGGGINNSVALG   192
>GCA_000455945.1    NG-QYQYKNEGNNSYMIGNKNKIASGSDDNFILGNNVHIGGGINNSVALG   205
>GCA_000178895.1    NG-QYQYKNEGNNSYMIGNKNKIASGSNDNFILGNNVHIGGGINNSVALG   205
>GCA_001510735.1    NG-QYQYKNEGNNSYMIGNKNKIASGSNDNFILGNNVHIGGGINNSVALG   205
>GCA_001296165.1    NG-QYQYKNEGNNSYMIGNKNKIASGSDDNFILGNNVHIGGGINNSVALG   205
>GCA_001296185.1    NG-QYQYKNEGNNSYMIGNKNKIASGSNDNFILGNNVHIGGGINNSVALG   205
>GCA_002211605.1    NG-QYQYKNEGNNSYMIGNKNKIASGSDDNFILGNNVHIGGGINNSVALG   205
>R24394             NG-QYQYKNEGNNSYMIGNKNKIASGSNDNFILGNNVHIGGGINNSVALG   205
>GCA_000158255.2    NG-SYQYKNEGNNSYMIGNKNKIASGSNDNFILGNNVHIGGGINNSVALG   177
>GCA_000162235.2    NG-SYQYKNEGNNSYMIGNKNKIASGSDDNFILGNNVHIGGGINNSVALG   177
>GCA_000182945.1    NG-SYQYKNEGNNSYMIGNKNKIASGSNDNFILGNNVHIGGGINNSVALG   177
>GCA_000347315.1    NG-SYQYKNEGNNSYMIGNKNKIASGSDDNFILGNNVHIGGGINNSVALG   177
>GCA_000479205.1    NG-QHLYKNEGNNSYMIGNKNKIASGSDDNFILGNNVHIGGGINNSVALG   177
>GCA_000517705.1    NG-QHLYKNEGNNSYMIGNKNKIASGSDDNFILGNNVHIGGGINNSVALG   177
>GCA_001296125.1    NG-SYQYKNEGNNSYMIGNKNKIASGSNDNFILGNNVHIGGGINNSVALG   177
>GCA_001810995.1    NG-SYQYKNEGNNSYMIGNKNKIASGSNDNFILGNNVHIGGGINNSVALG   177
>GCA_001854465.1    NG-QHLYKNEGNNSYMIGNKNKIASGSDDNFILGNNVHIGGGINNSVALG   177
>GCA_002749995.1    NG-SYQYKNEGNNSYMIGNKNKIASGSNDNFILGNNVHIGGGINNSVALG   177
>GCA_002764055.1    NG-SYQYKNEGNNSYMIGNKNKIASGSNDNFILGNNVHIGGGINNSVALG   177
>GCA_000455965.1    NG-SYQYKNEGNNSYMIGNKNKIASGSNDNFILGNNVHIGGGINNSVALG   177
>R26872             NG-SYQYKNEGNNSYMIGNKNKIASGSDDNFILGNNVHIGGGINNSVALG   177
>R28385             NG-SYQYKNEGNNSYMIGNKNKIASGSNDNFILGNNVHIGGGINNSVALG   177
>R28400             NG-SYQYKNEGNNSYMIGNKNKIASGSDDNFILGNNVHIGGGINNSVALG   177
>2B17               NG-SYQYKNEGNNSYMIGNKNKIASGSNDNFILGNNVHIGGGINNSVALG   177
>R29976             NG-SYQYKNEGNNSYMIGNKNKIASGSNDNFILGNNVHIGGGINNSVALG   177
>R30464             NG-QHLYKNEGNNSYMIGNKNKIASGSDDNFILGNNVHIGGGINNSVALG   177
>R30604             NG-SYQYKNEGNNSYMIGNKNKIASGSDDNFILGNNVHIGGGINNSVALG   177
>R31249             NG-SYQYKNEGNNSYMIGNKNKIASGSNDNFILGNNVHIGGGINNSVALG   177
>R32310             NG-SYQYKNEGNNSYMIGNKNKIASGSNDNFILGNNVHIGGGINNSVALG   177
>R33458             NG-SYQYKNEGNNSYMIGNKNKIASGSNDNFILGNNVHIGGGINNSVALG   177
>R33533             NG-SYQYKNEGNNSYMIGNKNKIASGSNDNFILGNNVHIGGGINNSVALG   177
>GCA_002243405.1    NG-SYQYKNEGNNSYMIGNKNKIASGSDDNFILGNNVHIGGGINNSVALG   177
CbpFb
>2B16               N-GQYQYKNEGNNSYMIGNKNKIASGSDDNFILGNNVTIGAGVQKSVVLG   261
>2B2                N-GQYKYKNEGNNSYMIGNYNKIASGSDDNFILGNNVTIGAGVQKSVVLG   261
>2B3                NSGSHLYKNEGNNSYMIGNKNKIASGSDDNFILGNNVEIGAGVQKSVVLG   262
>2B4                N-GQYKYKNEGNNSYMIGNYNKIASGSDDNFILGNNVTIGAGVQKSVVLG   261
>R5001              N-GQYKYKNEGNNSYMIGNYNKIAKDSNDNFILGNNVEIGAGVQKSVVLG   248
>GCA_001546435.1    DTGKYIYKNEGNNSYMIGNKNKIASGSNDNFILGNNVEIGAGVQKSVVLG   204
>R15792             N-GQYKNKNEGNNSYMIGNYNKIAKDSNDNFILGNNVEIGAGVQKSVVLG   191
CbpFc
>R16531_1           NG-QYQYKNEGHNSYMIGNKNKIAKGSNDNFILGNNVSIGKGIQNSVALG   177
```

```
                    301        311        321        331        341
CbpFa
>GCA_000007325.1    NNSTVSASNTVSVGSSTLKRKIVNVGDGAISANSSDAVTGRQLYSGNGID   242
>GCA_000163915.2    NNSTVSASNTVSVGSSTLKRKIVNVGDGEISASSTDAVTGRQLYSGNGID   242
>R18528             NNSTVSASNTVSVGSSTLKRKIVNVGDGAISANSSDAVTGRQLYSGNGID   242
>R28211             NNSTVSASNTVSVGSSTLKRKIVNVGDGEISASSTDAVTGRQLYSGNGID   242
>R32935             NNSTVSASNTVSVGSSTLKRKIVNVGDGAISANSSDAVTGRQLYSGNGID   242
>GCA_000455945.1    NNSTVSASNTVSVGSSTLKRKIVNVGDGAISANSSDAVTGRQLYSGNGID   255
>GCA_000178895.1    NNSTVSASNTVSVGSSTLKRKIVNVGDGAISANSSDAVTGRQLYSGNGID   255
>GCA_001510735.1    NNSTVSASNTVSVGSSTLKRKIVNVGDGAISANSSDAVTGRQLYSGNGID   255
>GCA_001296165.1    NNSTVSASNTVSVGSSTLKRKIVNVGDGAISANSSDAVTGRQLYSGNGID   255
>GCA_001296185.1    NNSTVSASNTVSVGSSTLKRKIVNVGDGAISANSSDAVTGRQLYSGNGID   255
>GCA_002211605.1    NNSTVSASNTVSVGSSTLKRKIVNVGDGAISANSSDAVTGRQLYSGNGID   255
>R24394             NNSTVSASNTVSVGSSTLKRKIVNVGDGAISANSSDAVTGRQLYSGNGID   255
>GCA_000158255.2    NNSTVSASNTVSVGSSTLKRKIVNVGDGEISASSTDAVTGRQLYSGNGID   227
>GCA_000162235.2    NNSTVSASNTVSVGSSTLKRKIVNVGDGEISASSTDAVTGRQLYSGNGID   227
>GCA_000182945.1    NNSTVSASNTVSVGSSTLKRKIVNVGDGEISASSTDAVTGRQLYSGNGID   227
>GCA_000347315.1    NNSTVSASNTVSVGSSTLKRKIVNVGDGEISASSTDAVTGRQLYSGNGID   227
>GCA_000479205.1    NNSTVSASNTVSVGSSTLKRKIVNVGDGEISASSTDAVTGRQLYSGNGID   227
>GCA_000517705.1    NNSTVSASNTVSVGSSTLKRKIVNVGDGEISASSTDAVTGRQLYSGNGID   227
>GCA_001296125.1    NNSTVSASNTVSVGSSTLKRKIVNVGDGEISASSTDAVTGRQLYSGNGID   227
>GCA_001810995.1    NNSTVSASNTVSVGSSTLKRKIVNVGDGEISASSTDAVTGRQLYSGNGID   227
>GCA_001854465.1    NNSTVSASNTVSVGSSTLKRKIVNVGDGEISASSTDAVTGRQLYSGNGID   227
>GCA_002749995.1    NNSTVSASNTVSVGSSTLKRKIVNVGDGEISASSTDAVTGRQLYSGNGID   227
>GCA_002764055.1    NNSTVSASNTVSVGSSTLKRKIVNVGDGEISASSTDAVTGRQLYSGNGID   227
>GCA_000455965.1    NNSTVSASNTVSVGSSTLKRKIVNVGDGEISASSTDAVTGRQLYSGNGID   227
>R26872             NNSTVSASNTVSVGSSTLKRKIVNVGDGTISANSSDAVTGRQLYSGNGID   227
>R28385             NNSTVSASNTVSVGSSTLKRKIVNVGDGEISASSTDAVTGRQLYSGNGID   227
>R28400             NNSTVSASNTVSVGSSTLKRKIVNVGDGEISASSTDAVTGRQLYSGNGID   227
>2B17               NNSTVSASNTVSVGSSTLKRKIVNVGDGEISASSTDAVTGRQLYSGNGID   227
>R29976             NNSTVSASNTVSVGSSTLKRKIVNVGDGEISASSTDAVTGRQLYSGNGID   227
>R30464             NNSTVSASNTVSVGSSTLKRKIVNVGDGEISASSTDAVTGRQLYSGNGID   227
>R30604             NNSTVSASNTVSVGSSTLKRKIVNVGDGEISASSTDAVTGRQLYSGNGID   227
>R31249             NNSTVSASNTVSVGSSTLKRKIVNVGDGEISASSTDAVTGRQLYSGNGID   227
>R32310             NNSTVSASNTVSVGSSTLKRKIVNVGDGEISASSTDAVTGRQLYSGNGID   227
>R33458             NNSTVSASNTVSVGSSTLKRKIVNVGDGEISASSTDAVTGRQLYSGNGID   227
>R33533             NNSTVSASNTVSVGSSTLKRKIVNVGDGEISASSTDAVTGRQLYSGNGID   227
>GCA_002243405.1    NNSTVSASNTVSVGSSTLKRKIVNVGDGEISASSTDAVTGRQLYSGNGID   227
CbpFb
>2B16               DGSASGGSNTVSVGSSTLQRKIVNVADGTISATSTDAVTGRQLYSGDGID   311
>2B2                DGSASGGSNTVSVGSSTLQRKIVNVADGTISATSTDAVTGRQLYSGDGID   311
>2B3                DGSASGGSNTVSVGSSTLQRKIVNVADGTISATSTDAVTGRQLYSGDGID   312
>2B4                DGSASGGSNTVSVGSSTLQRKIVNVADGTISATSTDAVTGRQLYSGDGID   311
>R5001              DGSASGGSNTVSVGSSTLQRKIVNVADGTISATSTDAVTGRQLYSGDGID   298
>GCA_001546435.1    DGSASGGSNTVSVGSSTLQRKIVNVADGTISATSTDAVTGRQLYSGDGID   254
>R15792             DGSASGGSNTVSVGSSTLQRKIANVADGTISATSTDAVTGRQLYSGDGID   241
CbpFc
>R16531_1           NNSTVTASNTVSVGSATLKRKIVNVGDGEVSATSSDAVTGKQLYRGEGID   227
```

```
                    351        361        371        381        391
CbpFa
>GCA_000007325.1    TAAWQNKLNVTRKNDYKDANDIDVNKWKAKLGVG--SGGGGGAPVDAYTK   290
>GCA_000163915.2    TAAWQNKLNVTKKNDYKDANDIDVNKWRTKLGVG--SGGSGGAPVDSYTK   290
>R18528             TAAWQNKLNVTRKNDYKDANDIDVNKWKAKLGVG--SGGGGGAPVDAYTK   290
>R28211             TAAWQNKLNVTKKNDYKDANDIDVNKWRTKLGVG--SGGVGGAPVDSYTK   290
>R32935             TAAWQNKLNVTRKNDYKDANDIDVNKWKAKLGVG--SGGGGGAPVDAYTK   290
>GCA_000455945.1    TAAWQNKLNVTRKNDYKDANDIDVNKWKAKLGVG--SGGGGGAPVDAYTK   303
>GCA_000178895.1    TAAWQNKLNVTRKNDYKDANDIDVNKWKAKLGVG--SGGGGGAPVDAYTK   303
>GCA_001510735.1    TAAWQNKLNVTRKNDYKDANDIDVNKWKAKLGVG--SGGGGGAPVDAYTK   303
>GCA_001296165.1    TAAWQNKLNVTRKNDYKDANDIDVNKWKAKLGVG--SGGGGGAPVDAYTK   303
>GCA_001296185.1    TAAWQNKLNVTRKNDYKDANDIDVNKWKAKLGVG--SGGGGGAPVDAYTK   303
>GCA_002211605.1    TAAWQNKLNVTRKNDYKDANDIDVNKWKAKLGVG--SGGGGGAPVDAYTK   303
>R24394             TAAWQNKLNVTRKNDYKDANDIDVNKWKAKLGVG--SGGGGGAPVDAYTK   303
>GCA_000158255.2    TAAWQNKLNVTKKNDYKDANDIDVNKWRTKLGVG--SGGGGGAPVDSYTK   275
>GCA_000162235.2    TAAWQNKLNVTRKNDYKDANDIDVNKWRTKLGVG--SGGGGGAPVDSYTK   275
>GCA_000182945.1    TAAWQNKLNVTKKNDYKDANDIDVNKWRTKLGVG--SGGGGGAPVDSYTK   275
>GCA_000347315.1    TAAWQNKLNVTRKNDYKDANDIDVNKWRTKLGVG--SGGGGGAPVDSYTK   275
>GCA_000479205.1    TAAWQNKLNVTKKNDYKDANDIDVNKWRTKLGVG--SGGGGGAPVDSYTK   275
>GCA_000517705.1    TAAWQNKLNVTKKNDYKDANDIDVNKWRTKLGVG--SGGGGGAPVDSYTK   275
>GCA_001296125.1    TAAWQNKLNVTKKNDYKDANDIDVNKWRTKLGVG--SGGGGGAPVDSYTK   275
>GCA_001810995.1    TAAWQNKLNVTKKNDYKDANDIDVNKWRTKLGVG--SGGGGGAPVDSYTK   275
>GCA_001854465.1    TAAWQNKLNVTKKNDYKDANDIDVNKWRTKLGVG--SGGGGGAPVDSYTK   275
>GCA_002749995.1    TAAWQNKLNVTKKNDYKDANDIDVNKWRTKLGVG--SGGGGGAPVDSYTK   275
>GCA_002764055.1    TAAWQNKLNVTKKNDYKDANDIDVNKWRTKLGVG--SGGGGGAPVDSYTK   275
>GCA_000455965.1    TAAWQNKLNVTKKNDYKDANDIDVNKWRTKLGVG--SGGGGGAPVDSYTK   275
>R26872             TAAWQNKLNVTKKNDYKDANDIDVNKWRTKLGVG--SGGGGGAPVDSYTK   275
>R28385             TAAWQNKLNVTKKNDYKDANDIDVNKWRTKLGVG--SGGGGGAPVDSYTK   275
>R28400             TAAWQSKLNVTKKNDYKDANDIDVNKWRTKLGVGG-SGGGGGAPVDSYTK   276
>2B17               TAAWQNKLNVTKKNDYKDANDIDVNKWRTKLGVG--SGGGGGAPVDSYTK   275
>R29976             TAAWQNKLNVTKKNDYKDANDIDVNKWRTKLGVG--SGGGGGAPVDSYTK   275
>R30464             TAAWQNKLNVTKKNDYKDANDIDVNKWRTKLGVG--SGGGGGAPVDSYTK   275
>R30604             TAAWQNKLNVTRKNDYKDANDIDVNKWRTKLGVG--SGGGGGAPVDSYTK   275
>R31249             TAAWQNKLNVTKKNDYKDANDIDVNKWRTKLGVG--SGGGGGAPVDSYTK   275
>R32310             TAAWQNKLNVTKKNDYKDANDIDVNKWRTKLGVG--SGGGGGAPVDSYTK   275
>R33458             TAAWQNKLNVTKKNDYKDANDIDVNKWRTKLGVG--SGGGGGAPVDSYTK   275
>R33533             TAAWQNKLNVTKKNDYKDANDIDVNKWRTKLGVG--SGGGGGAPVDSYTK   275
>GCA_002243405.1    TAAWQNKLNVTRKNDYKDANDIDVNKWRTKLGVG--SGGGGGAPVDAYTK   275
CbpFb
>2B16               ----------------------VNKWRTKLGVSSGGGASGGAPGDAYTK   338
>2B2                ----------------------VNKWRTKLGVSSGGGASGGAPGDAYTK   338
>2B3                ----------------------VNKWRTKLGVSSGGGASGGAPGDAYTK   339
>2B4                ----------------------VNKWRTKLGVSSGGGASGGAPGDAYTK   338
>R5001              ----------------------VNKWRTRLGVGSGGGAGGGAPVDAYTK   325
>GCA_001546435.1    ----------------------VNKWRTRLGVGSGGGAGGGAPVDAYTK   281
>R15792             ----------------------VNKWRTRLGVGSGGGAGGGAPVDAYTK   268
CbpFc
>R16531_1           ----------------------VNAWRAKLGVG----------------   238
```

```
                    401       411       421       431       441
CbpFa
>GCA_000007325.1    SEADNKFANKTDLNDYTKKDDYKDANGIDVDKWKAKLGTGAGTADIENLR   340
>GCA_000163915.2    SEADNKFANKTDLDNYTKKDDYKDANGIDVDKWKAKLGTGAGTADIENLR   340
>R18528             SEADNKFANKTDLNDYTKKDDYKDANGIDVDKWKAKLGTGAGTADIENLR   340
>R28211             SEADNKFANKTDLDNYTKKDDYKDANGIDVDKWKAKLGTGAGTADIENLR   340
>R32935             SEADNKFANKTDLNDYTKKDDYKDANGIDVDKWKAKLGTGAGTADIENLR   340
>GCA_000455945.1    SEADNKFANKTDLNDYTKKDDYKDANGIDVDKWKAKLGTGAGTADIENLR   353
>GCA_000178895.1    SEADNKFANKTDLNDYTKKDDYKDANGIDVDKWKAKLGTGAGTADIENLR   353
>GCA_001510735.1    SEADNKFANKTDLNDYTKKDDYKDANGIDVDKWKAKLGTGAGTADIENLR   353
>GCA_001296165.1    SEADNKFANKTDLNDYTKKDDYKDANGIDVDKWKAKLGTGAGTADIENLR   353
>GCA_001296185.1    SEADNKFANKTDLNDYTKKDDYKDANGIDVDKWKAKLGTGAGTADIENLR   353
>GCA_002211605.1    SEADNKFANKTDLNDYTKKDDYKDANGIDVDKWKAKLGTGAGTADIENLR   353
>R24394             SEADNKFANKTDLNDYTKKDDYKDANGIDVDKWKAKLGTGAGTADIENLR   353
>GCA_000158255.2    SEADNKFANKTDLDNYTKKDDYKDANGIDVDKWKAKLGTGAGTADIENLR   325
>GCA_000162235.2    SEADNKFANKTDLDNYTKKDDYKDANGIDVDKWKAKLGTGAGTADIENLR   325
>GCA_000182945.1    SEADNKFANKTDLDNYTKKDDYKDANGIDVDKWKAKLGTGAGTADIENLR   325
>GCA_000347315.1    SEADNKFANKTDLDNYTKKDDYKDANGIDVDKWKAKLGTGAGTADIENLR   325
>GCA_000479205.1    SEADNKFANKTDLDNYTKKDDYKDANGIDVDKWKAKLGTGAGTADIENLR   325
>GCA_000517705.1    SEADNKFANKTDLDNYTKKDDYKDANGIDVDKWKAKLGTGAGTADIENLR   325
>GCA_001296125.1    SEADNKFANKTDLDNYTKKDDYKDANGIDVDKWKAKLGTGAGTADIENLR   325
>GCA_001810995.1    SEADNKFANKTDLDNYTKKDDYKDANGIDVDKWKAKLGTGAGTADIENLR   325
>GCA_001854465.1    SEADNKFANKTDLDNYTKKDDYKDANGIDVDKWKAKLGTGAGTADIENLR   325
>GCA_002749995.1    SEADNKFANKTDLDNYTKKDDYKDANGIDVDKWKAKLGTGAGTADIENLR   325
>GCA_002764055.1    SEADNKFANKTDLDNYTKKDDYKDANGIDVDKWKAKLGTGAGTADIENLR   325
>GCA_000455965.1    SEADNKFANKTDLDNYTKKDDYKDANGIDVDKWKAKLGTGAGTADIENLR   325
>R26872             SEADNKFANKTDLDNYTKKDDYKDANGIDIDKWKAKLGTGAGTADIENLR   325
>R28385             SEADNKFANKTDLDNYTKKDDYKDANGIDVDKWKAKLGTGAGTADIENLR   325
>R28400             SEADNKFANKTDLDNYTKKDDYKDANGIDVDKWKAKLGTGAGTADIENLR   326
>2B17               SEADNKFANKTDLDNYTKKDDYKDANGIDVDKWKAKLGTGAGTADIENLR   325
>R29976             SEADNKFANKTDLDNYTKKDDYKDANGIDVDKWKAKLGTGAGTADIENLR   325
>R30464             SEADNKFANKTDLDNYTKKDDYKDANGIDVDKWKAKLGTGAGTADIENLR   325
>R30604             SEADNKFANKTDLDNYTKKDDYKDANGIDVDKWKAKLGTGAGTADIENLR   325
>R31249             SEADNKFANKTDLDNYTKKDDYKDANGIDVDKWKAKLGTGAGTADIENLR   325
>R32310             SEADNKFANKTDLDNYTKKDDYKDANGIDVDKWKAKLGTGAGTADIENLR   325
>R33458             SEADNKFANKTDLDNYTKKDDYKDANGIDVDKWKAKLGTGAGTADIENLR   325
>R33533             SEADNKFANKTDLDNYTKKDDYKDANGIDVDKWKAKLGTGAGTADIENLR   325
>GCA_002243405.1    SEADNKFANKTDLDNYTKKDDYKDANGIDVDKWKAKLGT--GTANIENLR   323
CbpFb
>2B16               SEADNKFTSK---------DDYKDANGIDVDKWKAKLGTGGSSSDIQNLR   379
>2B2                SEADNKFTSK---------DDYRDANGIDVDKWKAKLGTGASSTDIQNLR   379
>2B3                SEADNKFTSK---------DDYKDANGIDVDKWKAKLGTGGGSADIQNLR   380
>2B4                SEADNKFTSK---------DDYRDANGIDVDKWKAKLGTGASSTDIQNLR   379
>R5001              SEADNKFANKTDLDNYTKKDDYKDANGIDVDKWKAKLGTGADSADIQNLR   375
>GCA_001546435.1    SEADNKFANKTDLDNYTKKDDYKDANGIDVDKWKAKLGTGAGSADIQNLR   331
>R15792             SEADNKFANKTDLDNYTKKDDYKDANGIDVDKWKAKLGTGADSADIQNLR   318
CbpFc
>R16531_1           ------------------------------------TGS--ADIQNLR   248
```

```
                    451        461        471        481        491
CbpFa
>GCA_000007325.1    NEVNEKIDDVKDEVRTVGSLSAALAGLHPMQYDPKAPVQVMAALGHYRDK   390
>GCA_000163915.2    NEVNEKIDDVKDEVRTVGSLSAALAGLHPMQYDPKAPVQVMAALGHYRDK   390
>R18528             NEVNEKIDNVKDEVRTVGSLSAALAGLHPMQYDPKAPVQVMAALGHYKDK   390
>R28211             NEVNEKIDDVKDEVRTVGSLSAALAGLHPMQYDPKAPVQVMAALGHYRDK   390
>R32935             NEVNEKIDDVKDEVRTVGSLSAALAGLHPMQYDPKAPVQVMAALGHYRDK   390
>GCA_000455945.1    NEVNEKIDNVKDEVRTVGSLSAALAGLHPMQYDPKAPVQVMAALGHYRDK   403
>GCA_000178895.1    NEVNEKIDDVKDEVRTVGSLSAALAGLHPMQYDPKAPVQVMAALGHYRDK   403
>GCA_001510735.1    NEVNEKIDNVKDEVRTVGSLSAALAGLHPMQYDPKAPVQVMAALGHYRDK   403
>GCA_001296165.1    NEVNEKIDNVKDEVRTVGSLSAALAGLHPMQYDPKAPVQVMAALGHYRDK   403
>GCA_001296185.1    NEVNEKIDNVKDEVRTVGSLSAALAGLHPMQYDPKAPVQVMAALGHYRDK   403
>GCA_002211605.1    NEVNEKIDNVKDEVRTVGSLSAALAGLHPMQYDPKAPVQVMAALGHYRDK   403
>R24394             NEVNEKIDDVKDEVRTVGSLSAALAGLHPMQYDPKAPVQVMAALGHYRDK   403
>GCA_000158255.2    NEVNEKIDDVEDEVRTVGSLSAALAGLHPMQYDPKAPVQVMAALGHYRDK   375
>GCA_000162235.2    NEVNEKIDDVKDEVRTVGSLSAALAGLHPMQYDPKAPVQVMAALGHYRDK   375
>GCA_000182945.1    NEVNEKIDDVKDEVRTVGSLSAALAGLHPMQYDPKAPVQVMAALGHYRDK   375
>GCA_000347315.1    NEVNEKIDDVKDEVRTVGSLSAALAGLHPMQYDPKAPVQVMAALGHYRDK   375
>GCA_000479205.1    NEVNEKIDDVKDEVRTVGSLSAALAGLHPMQYDPKAPVQVMAALGHYRDK   375
>GCA_000517705.1    NEVNEKIDDVKDEVRTVGSLSAALAGLHPMQYDPKAPVQVMAALGHYRDK   375
>GCA_001296125.1    NEVNEKIDDVEDEVRTVGSLSAALAGLHPMQYDPKAPVQVMAALGHYRDK   375
>GCA_001810995.1    NEVNEKIDDVKDEVRTVGSLSAALAGLHPMQYDPKAPVQVMAALGHYRDK   375
>GCA_001854465.1    NEVNEKIDDVKDEVRTVGSLSAALAGLHPMQYDPKAPVQVMAALGHYRDK   375
>GCA_002749995.1    NEVNEKIDDVEDEVRTVGSLSAALAGLHPMQYDPKAPVQVMAALGHYRDK   375
>GCA_002764055.1    NEVNEKIDDVKDEVRTVGSLSAALAGLHPMQYDPKAPVQVMAALGHYRDK   375
>GCA_000455965.1    NEVNEKIDDVEDEVRTVGSLSAALAGLHPMQYDPKAPVQVMAALGHYRDK   375
>R26872             NEVNEKIDDVKDEVRTVGSLSAALAGLHPMQYDPKAPVQVMAALGHYRDK   375
>R28385             NEVNEKIDDVKDEVRTVGSLSAALAGLHPMQYDPKAPIQVMAALGHYRDK   375
>R28400             NEVNEKIDDVKDEVRTVGSLSAALAGLHPMQYDPKAPVQVMAALGHYRDK   376
>2B17               NEVNEKIDDVKDEVRTVGSLSAALAGLHPMQYDPKAPVQVMAALGHYRDK   375
>R29976             NEVNEKIDDVKDEVRTVGSLSAALAGLHPMQYDPKAPVQVMAALGHYRDK   375
>R30464             NEVNEKIDDVKDEVRTVGSLSAALAGLHPMQYDPKAPVQVMAALGHYRDK   375
>R30604             NEVNEKIDDVKDEVRTVGSLSAALAGLHPMQYDPKAPVQVMAALGHYRDK   375
>R31249             NEVNEKIDDVKDEVRTVGSLSAALAGLHPMQYDPKAPVQVMAALGHYRDK   375
>R32310             NEVNEKIDDVKDEVRTVGSLSAALAGLHPMQYDPKAPVQVMAALGHYRDK   375
>R33458             NEVNEKIDDVEDEVRTVGSLSAALAGLHPMQYDPKAPVQVMAALGHYRDK   375
>R33533             NEVNEKIDDVKDEVRTVGSLSAALAGLHPMQYDPKAPVQVMAALGHYRDK   375
>GCA_002243405.1    NEVNEKIDNVKDEVRTVGSLSAALAGLHPMQYDPKAPVQVMAALGHYKDK   373
CbpFb
>2B16               NEVNEKIDNVKDEVRGVGSLSAALAGLHPMQYDPKAPAQVMAALGHYKNR   429
>2B2                NEVNEKIDNVKDEVRGVGSLSAALAGLHPMQYDPKAPAQVMAALGHYKDR   429
>2B3                NEVNEKIDNVKDEVRGVGSLSAALAGLHPMQYDPKAPAQVMAALGHYKDR   430
>2B4                NEVNEKIDNVKDEVRGVGSLSAALAGLHPMQYDPKAPAQVMAALGHYKDR   429
>R5001              NEVYERIDNVKDEVRDVGSLSAALAGLHPMQYDPKAPAQVMAALGHYKDR   425
>GCA_001546435.1    NEVYERIDNVKDEVRDVGSLSAALAGLHPMQYDPKAPAQVMAALGHYKDR   381
>R15792             NEVYERIDNVKDEVRDVGSLSAALAGLHPMQYDPKAPAQVMAALGHYKDR   368
CbpFc
>R16531_1           NEVYERIDNVKDEVRDVGSLSAALAGLHPMQYDPKAPAQVMAALGHYKDR   298
```

```
                    501       511       521       531       541
CbpFa
>GCA_000007325.1    QSVAVGASYYFNDRFMMSTGIALSGEKRTKTMANVGFTLKLGKGSGVTYD   440
>GCA_000163915.2    QSVAVGASYYFNDRFMMSTGIALSGEKRTKTMANVGFTLKLGKGSGVAYN   440
>R18528             QSVAVGASYYFNDRFMMSTGIALSGEKRTETMANVGFTLKLGKGSGVTYD   440
>R28211             QSVAVGASYYFNDRFMMSTGIALSGEKRTKTMANVGFTLKLGKGSGVAYN   440
>R32935             QSVAVGASYYFNDRFMMSTGIALSGEKRTKTMANVGFTLKLGKGSGVTYD   440
>GCA_000455945.1    QSVAVGASYYFNDRFMMSTGIALSGEKRTKTMANVGFTLKLGKGSGVTYD   453
>GCA_000178895.1    QSVAVGASYYFNDRFMMSTGIALSGEKKTKTMANVGFTLKLGKGSGVTYD   453
>GCA_001510735.1    QSVAVGASYYFNDRFMMSTGIALSGEKRTKTMANVGFTLKLGKGSGVTYD   453
>GCA_001296165.1    QSVAVGASYYFNDRFMMSTGIALSGEKRTKTMANVGFTLKLGKGSGVTYD   453
>GCA_001296185.1    QSVAVGASYYFNDRFMMSTGIALSGEKRTKTMANVGFTLKLGKGSGVTYD   453
>GCA_002211605.1    QSVAVGASYYFNDRFMMSTGIALSGEKRTKTMANVGFTLKLGKGSGVTYD   453
>R24394             QSVAVGASYYFNDRFMMSTGIALSGEKKTKTMANVGFTLKLGKGSGVTYD   453
>GCA_000158255.2    QSVAVGASYYFNDRFMMSTGIALSGEKRTKTMANVGFTLKLGKGSGVAYN   425
>GCA_000162235.2    QSVAVGASYYFNDRFMMSTGIALSGEKRTKTMANVGFTLKLGKGSGVAYN   425
>GCA_000182945.1    QSVAVGASYYFNDRFMMSTGIALSGEKRTKTMANVGFTLKLGKGSGVTYN   425
>GCA_000347315.1    QSVAVGASYYFNDRFMMSTGIALSGEKRTKTMANVGFTLKLGKGSGVAYN   425
>GCA_000479205.1    QSVAVGASYYFNDRFMMSTGIALSGEKRTKTMANVGFTLKLGKGSGVAYN   425
>GCA_000517705.1    QSVAVGASYYFNDRFMMSTGIALSGEKRTKTMANVGFTLKLGKGSGVAYN   425
>GCA_001296125.1    QSVAVGASYYFNDRFMMSTGIALSGEKRTKTMANVGFTLKLGKGSGVAYN   425
>GCA_001810995.1    QSVAVGASYYFNDRFMMSTGIALSGEKRTKTMANVGFTLKLGKGSGVTYN   425
>GCA_001854465.1    QSVAVGASYYFNDRFMMSTGIALSGEKRTKTMANVGFTLKLGKGSGVAYN   425
>GCA_002749995.1    QSVAVGASYYFNDRFMMSTGIALSGEKRTKTMANVGFTLKLGKGSGVAYN   425
>GCA_002764055.1    QSVAVGASYYFNDRFMMSTGIALSGEKRTKTMANVGFTLKLGKGSGVTYN   425
>GCA_000455965.1    QSVAVGASYYFNDRFMMSTGIALSGEKRTKTMANVGFTLKLGKGSGVAYN   425
>R26872             QSVAVGASYYFNDRFMMSTGIALSGEKRTKTMANVGFTLKLGKGSGVTYN   425
>R28385             QSVAVGASYYFNDRFMMSTGIALSGEKRTKTMANVGFTLKLGKGSGVTYN   425
>R28400             QAVAVGASYYFNDRFMMSTGIALSGEKRTKTMANVGFTLKLGKGSGVTYN   426
>2B17               QSVAVGASYYFNDRFMMSTGIALSGEKRTKTMANVGFTLKLGKGSGVTYN   425
>R29976             QSVAVGASYYFNDRFMMSTGIALSGEKRTKTMANVGFTLKLGKGSGVAYN   425
>R30464             QSVAVGASYYFNDRFMMSTGIALSGEKRTKTMANVGFTLKLGKGSGVAYN   425
>R30604             QSVAVGASYYFNDRFMMSTGIALSGEKRTKTMANVGFTLKLGKGSGVAYN   425
>R31249             QSVAVGASYYFNDRFMMSTGIALSGEKRTKTMANVGFTLKLGKGSGVTYN   425
>R32310             QSVAVGASYYFNDRFMMSTGIALSGEKRTKTMANVGFTLKLGKGSGVTYN   425
>R33458             QSVAVGASYYFNDRFMMSTGIALSGEKRTKTMANVGFTLKLGKGSGVAYN   425
>R33533             QSVAVGASYYFNDRFMMSTGIALSGEKRTKTMANVGFTLKLGKGSGVAYN   425
>GCA_002243405.1    QSVAVGASYYFNDRFMMSTGIALSGEKRTETMANVGFTLKLGKGSGVTYN   423
CbpFb
>2B16               QAVAVGASYYFNDKFMMSTGVALSGEKRTEAMANVGFTLKIGKGSGTTYT   479
>2B2                QAVAVGASYYFNDKFMMSTGVALSGEKRTEAMANVGFTLKIGKGSGTTYT   479
>2B3                QAVAVGASYYFNDKFMMSTGVALSGEKRTEAMANVGFTLKIGKGSGTTYT   480
>2B4                QAVAVGASYYFNDKFMMSTGVALSGEKRTEAMANVGFTLKIGKGSGTTYT   479
>R5001              QAVAVGASYYFNDRFMMSTGVALSGEKKTKTMANVGFTLKLGKGSGTTYS   475
>GCA_001546435.1    QAVAVGASYYFNDRFMMSTGVALSGEKKTKTMANVGFTLKLGKGSGTTYS   431
>R15792             QAVAVGASYYFNDRFMMSTGVALSGEKKTKTMANVGFTLKLGKGSGTTYS   418
CbpFc
>R16531_1           QAVAVGASYYFNDRFMMSTGVALSGEKKTKTMANVGFTLKLGKGSGTTYS   348
```

```
                      551        561        571        581        591
CbpFa
>GCA_000007325.1      ETPQYVVQNEVKRLTVEN-------QELKERVRNLEEKLNMLLKNK        483
>GCA_000163915.2      ETPLYTIQDEVKRLTVENNKQAKENQELKERVRNLEEKLNMLLKNK        490
>R18528               ETPQYVVQNEVKRLTVEN-------QELKERLRNLEEKLETLLKNK        483
>R28211               ETPLYTIQDEVKRLTVENNKQAKENQELKERVRNLEEKLNMLLKNK        490
>R32935               ETPQYVVQNEVKRLTVEN-------QELKERVRNLEEKLNMLLKNK        483
>GCA_000455945.1      ETPQYVVQNEVKRLTVEN-------QELKERLRNLEEKLEMLLKNK        496
>GCA_000178895.1      ETPQYVVQNEVKRLTVEN-------QELKERVRNLEEKLNMLLKNK        496
>GCA_001510735.1      ETPQYVVQNEVKRLTVEN-------QELKERLRNLEEKLEMLLKNK        496
>GCA_001296165.1      ETPQYVVQNEVKRLTVEN-------QELKERLRNLEEKLEMLLKNK        496
>GCA_001296185.1      ETPQYVVQNEVKRLTVEN-------QELKERLRNLEEKLEMLLKNK        496
>GCA_002211605.1      ETPQYVVQNEVKRLTVEN-------QELKERLRNLEEKLEMLLKNK        496
>R24394               ETPQYVVQNEVKRLTVEN-------QELKERVRNLEEKLNMLLKNK        496
>GCA_000158255.2      ETPLYTIQDEVKRLTVENNKQAKENQELKERVRNLEEKLNMLLKNK        475
>GCA_000162235.2      ETPLYTIQDEVKRLTVENNKQAKENQELKERVRNLEEKLNMLLKNK        475
>GCA_000182945.1      ETPLYTIQDEVKRLTVENNKQAKENQELKERVRNLEEKLNMLLKNK        475
>GCA_000347315.1      ETPLYTIQDEVKRLTVENNKQAKENQELKERVRNLEEKLNMLLKNK        475
>GCA_000479205.1      ETPLYTIQDEVKRLTVENNKQAKENQELKERVRNLEEKLNMLLKNK        475
>GCA_000517705.1      ETPLYTIQDEVKRLTVENNKQAKENQELKERVRNLEEKLNMLLKNK        475
>GCA_001296125.1      ETPLYTIQDEVKRLTVENNKQAKENQELKERVRNLEEKLNMLLKNK        475
>GCA_001810995.1      ETPLYTIQDEVKRLTVENNKQAKENQELKERVRNLEEKLNMLLKNK        475
>GCA_001854465.1      ETPLYTIQDEVKRLTVENNKQAKENQELKERVRNLEEKLNMLLKNK        475
>GCA_002749995.1      ETPLYTIQDEVKRLTVENNKQAKENQELKERVRNLEEKLNMLLKNK        475
>GCA_002764055.1      ETPLYTIQDEVKRLTVENNKQAKENQELKERVRNLEEKLNMLLKNK        475
>GCA_000455965.1      ETPLYTIQDEVKRLTVENNKQAKENQELKERVRNLEEKLNMLLKNK        475
>R26872               ETPLYTIQDEVKRLTVENNKQAKENQELKERVRNLEEKLNMLLKNK        475
>R28385               ETPLYTIQDEVKRLTVENNKQAKENQELKERVRNLEEKLNMLLKNK        475
>R28400               ETPLYTIQDEVKRLTVENNKQAKENQELKERVRNLEEKLNMLLKNK        476
>2B17                 ETPLYTIQDEVKRLTVENNKQAKENQELKERVRNLEEKLNMLLKNK        475
>R29976               ETPLYTIQDEVKRLTIENNKQAKENQELKERVRNLEEKLNMLLKNK        475
>R30464               ETPLYTIQDEVKRLTVENNKQAKENQELKERVRNLEEKLNMLLKNK        475
>R30604               ETPLYTIQDEVKRLTVENNKQAKENQELKERVRNLEEKLNMLLKNK        475
>R31249               ETPLYTIQDEVKRLTVENNKQAKENQELKERVRNLEEKLNMLLKNK        475
>R32310               ETPLYTIQDEVKRLTVENNKQAKENQELKERVRNLEEKLNMLLKNK        475
>R33458               ETPLYTIQDEVKRLTVENNKQAKENQELKERVRNLEEKLNMLLKNK        475
>R33533               ETPLYTIQDEVKRLTVENNKQAKENQELKERVRNLEEKLNMLLKNK        475
>GCA_002243405.1      ETPQYVVQNEVKRLTVEN-------QELKERLRNLEEKLEMLLKNK        466
CbpFb
>2B16                 ETPQYVVQNEVKRLTVEN-------QELKERVRNLEEKLNMLLKNK        522
>2B2                  ETPQYVVQNEVKRLTVEN-------QELKERVRNLEEKLNMLLKNK        522
>2B3                  ETPQYVVQNEVKRLTVEN-------QELKERVRNLEEKLNMLLKNK        523
>2B4                  ETPQYVVQNEVKRLTVEN-------QELKERVRNLEEKLNMLLKNK        522
>R5001                ETPQYVVQNEVKRLTVEN-------QELKERLRNLEQKLEILLKNK        518
>GCA_001546435.1      ETPQYVVQNEVKRLTVEN-------QELKERLRNLEQKLEILLKNK        474
>R15792               ETPQYVVQNEVKRLTVEN-------QELKERLRNLEQKLEILLKNK        461
CbpFc
>R16531_1             ETPQYVVQNEVKRLTVEN-------QELKERVRNLEEKLNMLLKSK        391
```

Appendix E: List of *Fusobacterium* Reclassifications

Table S 2 | **List of all *Fusobacterium* classifications and reclassifications from CHAPTER 3.**

All *Fusobacterium* genomes listed on NCBI, were classified according to the MUMi, ANI and PLSA scores such that a consensus could be reached where every strain was within the species threshold. The proposed genus reclassification groups (A, B and C) are also shown – blanks are species that do not exist within the *Fusobacteriaceae* family or where there was ambiguity in the genome, such as for R33458.

Assembly Accession or WGS Strain ID	Organism Name (NCBI)	Strain ID	New Species Classification	Genus Group
GCA_000158275.2	*F. nucleatum* subsp. *animalis*	7_1	*animalis*	A
GCA_000158535.2	*F. nucleatum* subsp. *animalis*	D11	*animalis*	A
GCA_000162355.2	*F. nucleatum* subsp. *animalis*	3_1_33	*animalis*	A
GCA_000218645.2	*F. nucleatum* subsp. *animalis*	21_1A	*animalis*	A
GCA_000218655.1	*F. nucleatum* subsp. *animalis*	11_3_2	*animalis*	A
GCA_000220825.1	*F. nucleatum* subsp. *animalis*	ATCC 51191	*animalis*	A
GCA_000234075.2	*F. nucleatum* subsp. *polymorphum*	F0401	*animalis*	A
GCA_000242975.1	*F. nucleatum* subsp. *animalis*	OT 420	*animalis*	A
GCA_000273605.1	*F. nucleatum* subsp. *animalis*	F0419	*animalis*	A
GCA_000273625.1	*F. nucleatum* subsp. *animalis*	7_1	*animalis*	A
GCA_000400875.1	*F. nucleatum* subsp. *animalis*	4_8	*animalis*	A
GCA_000433695.1	*F.* sp.	CAG:649	*animalis*	A
GCA_000455985.1	*F. nucleatum* subsp. *animalis*	ChDC F324	*animalis*	A
GCA_000479225.1	*F. nucleatum*	CTI-5	*animalis*	A
GCA_000479245.1	*F. nucleatum*	CTI-3	*animalis*	A
GCA_000479285.1	*F. nucleatum*	CTI-1	*animalis*	A

GCA_000524215.1	*F.* sp.	CM1	*animalis*	A
GCA_001296085.1	*F. nucleatum* subsp. *animalis*	KCOM 1279	*animalis*	A
GCA_001296145.1	*F. nucleatum* subsp. *animalis*	KCOM 1325	*animalis*	A
GCA_001546435.1	*F. nucleatum*	MJR775 7B	*animalis*	A
GCA_001813745.1	*F.* sp.	HMSC06 5F01	*animalis*	A
GCA_002211645.1	*F. nucleatum* subsp. *animalis*	ChDC F332	*animalis*	A
GCA_002573475.1	*F. nucleatum* subsp. *animalis*	ChDC F318	*animalis*	A
GCA_002762005.1	*F. nucleatum* subsp. *animalis*	P2_CP	*animalis*	A
GCA_002762015.1	*F. nucleatum* subsp. *animalis*	P2_LM	*animalis*	A
GCA_002884895.1	*F. nucleatum*	UMB024 9	*animalis*	A
R15792		R15792	*animalis*	A
R18932		R18932	*animalis*	A
R30927		R30927	*animalis*	A
R5001		R5001	*animalis*	A
GCA_900015295.1	*F.* sp.		*closii* sp. nov.	B
GCA_000158235.1	*F. gonidiaformans*	3_1_5R	*gonidiaformans*	C
GCA_000158835.2	*F. gonidiaformans*	ATCC 25563	*gonidiaformans*	C
GCA_001546395.1	*F. equinum*	CMW839 6	*gonidiaformans*	C
GCA_003019695.1	*F. gonidiaformans*	ATCC 25563	*gonidiaformans*	C
GCA_000292935.1	*F. hwasookii*	ChDC F128	*hwasookii*	A
GCA_000455865.1	*F. hwasookii*	ChDC F145	*hwasookii*	A
GCA_000455885.1	*F. hwasookii*	ChDC F174	*hwasookii*	A
GCA_000455905.1	*F. hwasookii*	ChDC F206	*hwasookii*	A
GCA_000455925.1	*F. hwasookii*	ChDC F300	*hwasookii*	A

GCA_001455085.1	*F. hwasookii*	ChDC F206	*hwasookii*	A
GCA_001455105.1	*F. hwasookii*	ChDC F300	*hwasookii*	A
GCA_001455145.1	*F. hwasookii*	ChDC F174	*hwasookii*	A
GCA_900095705.1	*F. massiliense*	Marseille -P2749	*massiliense*	A
GCA_000158195.2	*F. mortiferum*	ATCC 9817	*mortiferum*	B
GCA_003019315.1	*F. mortiferum*	ATCC 9817	*mortiferum*	B
GCA_003438345.1	*F. mortiferum*	OM06-15BH	*mortiferum*	B
GCA_003014445.1	*F. naviforme*	ATCC 25832	*naviforme*	
GCA_900450945.1	*F. naviforme*	NCTC13 121	*naviforme*	
GCA_900450765.1	*F. necrogenes*	NCTC10 723	*necrogenes*	B
GCA_000158295.2	*F. necrophorum*	D12	*necrophorum*	C
GCA_000242215.1	*F. necrophorum* subsp. *funduliforme*	1_1_36S	*necrophorum*	C
GCA_000262225.1	*F. necrophorum* subsp. *funduliforme*	ATCC 51357	*necrophorum*	C
GCA_000292975.1	*F. necrophorum* subsp. *funduliforme*	Fnf 1007	*necrophorum*	C
GCA_000600355.1	*F. necrophorum* subsp. *funduliforme*	B35	*necrophorum*	C
GCA_000622045.1	*F. necrophorum*	HUN048	*necrophorum*	C
GCA_000691645.1	*F. necrophorum*	BL	*necrophorum*	C
GCA_000691665.1	*F. necrophorum*	DJ-1	*necrophorum*	C
GCA_000691685.1	*F. necrophorum*	BFTR-1	*necrophorum*	C
GCA_000691705.1	*F. necrophorum*	DAB	*necrophorum*	C
GCA_000691725.1	*F. necrophorum*	BFTR-2	*necrophorum*	C
GCA_000691745.1	*F. necrophorum*	DJ-2	*necrophorum*	C
GCA_000814775.1	*F. necrophorum* subsp. *funduliforme*	B35	*necrophorum*	C
GCA_001596475.1	*F. necrophorum* subsp. *funduliforme*	LS_1260	*necrophorum*	C

GCA_001596485.1	*F. necrophorum* subsp. *funduliforme*	LS_1264	*necrophorum*	C
GCA_001596495.1	*F. necrophorum* subsp. *funduliforme*	LS_1197	*necrophorum*	C
GCA_001597305.1	*F. necrophorum* subsp. *funduliforme*	LS_1195	*necrophorum*	C
GCA_001597315.1	*F. necrophorum* subsp. *funduliforme*	LS_1266	*necrophorum*	C
GCA_001597325.1	*F. necrophorum* subsp. *funduliforme*	LS_1280	*necrophorum*	C
GCA_001597335.1	*F. necrophorum* subsp. *funduliforme*	LS_1272	*necrophorum*	C
GCA_001597385.1	*F. necrophorum* subsp. *funduliforme*	LS_1291	*necrophorum*	C
GCA_001597395.1	*F. necrophorum* subsp. *funduliforme*	F1248	*necrophorum*	C
GCA_001597405.1	*F. necrophorum* subsp. *funduliforme*	F1285	*necrophorum*	C
GCA_001597445.1	*F. necrophorum* subsp. *funduliforme*	F1250	*necrophorum*	C
GCA_001597465.1	*F. necrophorum* subsp. *funduliforme*	F1267	*necrophorum*	C
GCA_001597475.1	*F. necrophorum* subsp. *funduliforme*	F1309	*necrophorum*	C
GCA_001597485.1	*F. necrophorum* subsp. *funduliforme*	F1314	*necrophorum*	C
GCA_001597525.1	*F. necrophorum* subsp. *funduliforme*	F1353	*necrophorum*	C
GCA_001597545.1	*F. necrophorum* subsp. *funduliforme*	F1330	*necrophorum*	C
GCA_001597565.1	*F. necrophorum* subsp. *funduliforme*	F1365	*necrophorum*	C
GCA_001597575.1	*F. necrophorum* subsp. *funduliforme*	F1351	*necrophorum*	C
GCA_002761995.1	*F. necrophorum* subsp. *funduliforme*	P1_CP	*necrophorum*	C
GCA_002762025.1	*F. necrophorum* subsp. *funduliforme*	P1_LM	*necrophorum*	C
GCA_003019715.1	*F. necrophorum* subsp. *funduliforme*	1_1_36S	*necrophorum*	C
GCA_900104395.1	*F. necrophorum*	ATCC 25286	*necrophorum*	C
GCA_900451075.1	*F. necrophorum* subsp. *necrophorum*	NCTC13 726	*necrophorum*	C
GCA_000007325.1	*F. nucleatum* subsp. *nucleatum*	ATCC 25586	*nucleatum*	A
GCA_000178895.1	*F. nucleatum* subsp. *nucleatum*	ATCC 23726	*nucleatum*	A

GCA_000455945.1	*F. nucleatum* subsp. *nucleatum*	ChDC F316	*nucleatum*	A
GCA_000479265.1	*F. nucleatum*	CTI-2	*nucleatum*	A
GCA_001296165.1	*F. nucleatum* subsp. *nucleatum*	KCOM 1322	*nucleatum*	A
GCA_001296185.1	*F. nucleatum* subsp. *nucleatum*	KCOM 1250	*nucleatum*	A
GCA_001510735.1	*F. nucleatum* subsp. *nucleatum*	ChDC F311	*nucleatum*	A
GCA_002211605.1	*F. nucleatum* subsp. *nucleatum*	ChDC F317	*nucleatum*	A
GCA_002243405.1	*F.* sp. *oral taxon 203*	W7671	*nucleatum*	A
GCA_003019295.1	*F. nucleatum* subsp. *nucleatum*	25586	*nucleatum*	A
GCA_003019785.1	*F. nucleatum* subsp. *nucleatum*	ATCC 23726	*nucleatum*	A
R18528		R18528	*nucleatum*	A
R24394		R24394	*nucleatum*	A
R28385		R28385	*nucleatum*	A
R28400		R28400	*nucleatum*	A
R32935		R32935	*nucleatum*	A
2B16		2B16	*oralis* sp. nov.	A
2B2		2B2	*oralis* sp. nov.	A
2B3		2B3	*oralis* sp. nov.	A
2B4		2B4	*oralis* sp. nov.	A
GCA_000235465.1	*F.* sp. *oral taxon 370*	F0437	*oralis* sp. nov.	A
R28427		R28427	*oralis* sp. nov.	A
R16531		R16531	*ovarium* sp. nov.	A
GCA_000622245.1	*F. perfoetens*	ATCC 29250	*perfoetens*	B
GCA_000158215.3	*F. periodonticum*	2_1_31	*periodonticum*	A
GCA_000163935.1	*F. periodonticum*	1_1_41FAA	*periodonticum*	A

GCA_000297655.1	*F. periodonticum*	D10	*periodonticum*	A
GCA_002761935.1	*F. periodonticum*	KCOM 1321	*periodonticum*	A
GCA_002761955.1	*F. periodonticum*	KCOM 1259	*periodonticum*	A
GCA_002763595.1	*F. periodonticum*	KCOM 1283	*periodonticum*	A
GCA_002763625.1	*F. periodonticum*	KCOM 1261	*periodonticum*	A
GCA_002763695.1	*F. periodonticum*	KCOM 1263	*periodonticum*	A
GCA_002763735.1	*F. periodonticum*	KCOM 1277	*periodonticum*	A
GCA_002763775.1	*F. periodonticum*	KCOM 1282	*periodonticum*	A
GCA_002763815.1	*F. periodonticum*	KCOM 2305	*periodonticum*	A
GCA_002763875.1	*F. periodonticum*	KCOM 2555	*periodonticum*	A
GCA_002763915.1	*F. periodonticum*	KCOM 1262	*periodonticum*	A
GCA_002763925.1	*F. periodonticum*	KCOM 2653	*periodonticum*	A
GCA_003019755.1	*F. periodonticum*	2_1_31	*periodonticum*	A
GCA_000153625.1	*F. nucleatum* subsp. *polymorphum*	ATCC 10953	*polymorphum*	A
GCA_000479185.1	*F. nucleatum*	CTI-6	*polymorphum*	A
GCA_000523555.1	*F. nucleatum*	13_3C	*polymorphum*	A
GCA_000524235.1	*F.* sp.	OBRC1	*polymorphum*	A
GCA_000524395.1	*F.* sp.	CM22	*polymorphum*	A
GCA_001433955.1	*F. nucleatum* subsp. *polymorphum*	ChDC F306	*polymorphum*	A
GCA_001455125.1	*F. nucleatum* subsp. *polymorphum*	ChDC F319	*polymorphum*	A
GCA_001457555.1	*F. nucleatum* subsp. *polymorphum*	NCTC10 562	*polymorphum*	A
GCA_001815715.1	*F.* sp.	HMSC06 4B11	*polymorphum*	A
GCA_002202115.1	*F. nucleatum* subsp. *polymorphum*	ChDC F218	*polymorphum*	A
GCA_002204435.1	*F. nucleatum* subsp. *polymorphum*	KCOM 1001	*polymorphum*	A

GCA_002211625.1	*F. nucleatum* subsp. *polymorphum*	KCOM 1275	*polymorphum*	A
GCA_002417615.1	*F. nucleatum*	12230 MIT 2016	*polymorphum*	A
GCA_002573625.1	*F. nucleatum* subsp. *polymorphum*	ChDC F37	*polymorphum*	A
GCA_002591465.1	*F. nucleatum* subsp. *polymorphum*	ChDC F309	*polymorphum*	A
GCA_002591475.1	*F. nucleatum* subsp. *polymorphum*	ChDC F290	*polymorphum*	A
GCA_002591505.1	*F. nucleatum* subsp. *polymorphum*	ChDC F113	*polymorphum*	A
GCA_002591515.1	*F. nucleatum* subsp. *polymorphum*	ChDC F305	*polymorphum*	A
GCA_002591545.1	*F. nucleatum* subsp. *polymorphum*	ChDC F186	*polymorphum*	A
GCA_002591555.1	*F. nucleatum* subsp. *polymorphum*	ChDC F175	*polymorphum*	A
GCA_002591585.1	*F. nucleatum* subsp. *polymorphum*	ChDC F313	*polymorphum*	A
GCA_002591645.1	*F. nucleatum* subsp. *polymorphum*	ChDC F330	*polymorphum*	A
GCA_002761915.1	*F. periodonticum*	KCOM 1265	*polymorphum*	A
GCA_003226385.1	*F. nucleatum*	12230	*polymorphum*	A
GCA_000160475.1	*F. periodonticum*	ATCC 33693	*pseudoperiodonticum* sp. nov.	A
GCA_002356455.1	*F. varium*	Fv113-g1	*pseudovarium* sp. nov.	B
GCA_000381725.1	*F. russii*	ATCC 25533 593A	*russii*	A
GCA_000158315.2	*F. ulcerans*	ATCC 49185	*ulcerans*	B
GCA_000242995.2	*F. ulcerans*	12_1B	*ulcerans*	B
GCA_003019675.1	*F. ulcerans*	ATCC 49185	*ulcerans*	B
GCA_900478315.1	*F. ulcerans*	NCTC12 112	*ulcerans*	B
GCA_000159915.2	*F. varium*	ATCC 27725	*varium*	B
GCA_001810475.1	*F.* sp.	HMSC07 3F01	*varium*	B
GCA_003019655.1	*F. varium*	ATCC 27725	*varium*	B

GCA_003436335.1	*F. varium*	TM07-10	*varium*	B
2B17		2B17	*vincentii*	A
GCA_000158255.2	*F. nucleatum* subsp. *vincentii*	4_1_13	*vincentii*	A
GCA_000162235.2	*F. nucleatum* subsp. *vincentii*	3_1_36A 2	*vincentii*	A
GCA_000163915.2	*F. nucleatum* subsp. *vincentii*	3_1_27	*vincentii*	A
GCA_000182945.1	*F. nucleatum* subsp. *vincentii*	ATCC 49256	*vincentii*	A
GCA_000279975.1	*F. nucleatum* subsp. *fusiforme*	ATCC 51190	*vincentii*	A
GCA_000347315.1	*F. nucleatum*	CC53	*vincentii*	A
GCA_000455965.1	*F. nucleatum* subsp. *vincentii*	ChDC F8	*vincentii*	A
GCA_000479205.1	*F. nucleatum*	CTI-7	*vincentii*	A
GCA_000517705.1	*F.* sp.	CM21	*vincentii*	A
GCA_001296125.1	*F. nucleatum* subsp. *vincentii*	KCOM 1231	*vincentii*	A
GCA_001810995.1	*F.* sp.	HMSC06 4B12	*vincentii*	A
GCA_001854465.1	*F. nucleatum*	AB1	*vincentii*	A
GCA_002749995.1	*F. nucleatum* subsp. *vincentii*	KCOM 2880	*vincentii*	A
GCA_002764055.1	*F. nucleatum* subsp. *vincentii*	KCOM 2931	*vincentii*	A
GCA_900450795.1	*F. nucleatum* subsp. *vincentii*	NCTC11 326	*vincentii*	A
R26872		R26872	*vincentii*	A
R28211		R28211	*vincentii*	A
R29976		R29976	*vincentii*	A
R30464		R30464	*vincentii*	A
R30604		R30604	*vincentii*	A
R31249		R31249	*vincentii*	A
R32310		R32310	*vincentii*	A

R33533		R33533	*vincentii*	A
GCA_000493815.1	*F. nucleatum* subsp. W1481	W1481	W1481	A
GCA_000437775.1	*F.* sp.	CAG:815		
GCA_000438175.1	*F.* sp.	CAG:439		
R33458		R33458		

Appendix F: CEACAM1 N-terminal Domain Molecular Dynamics

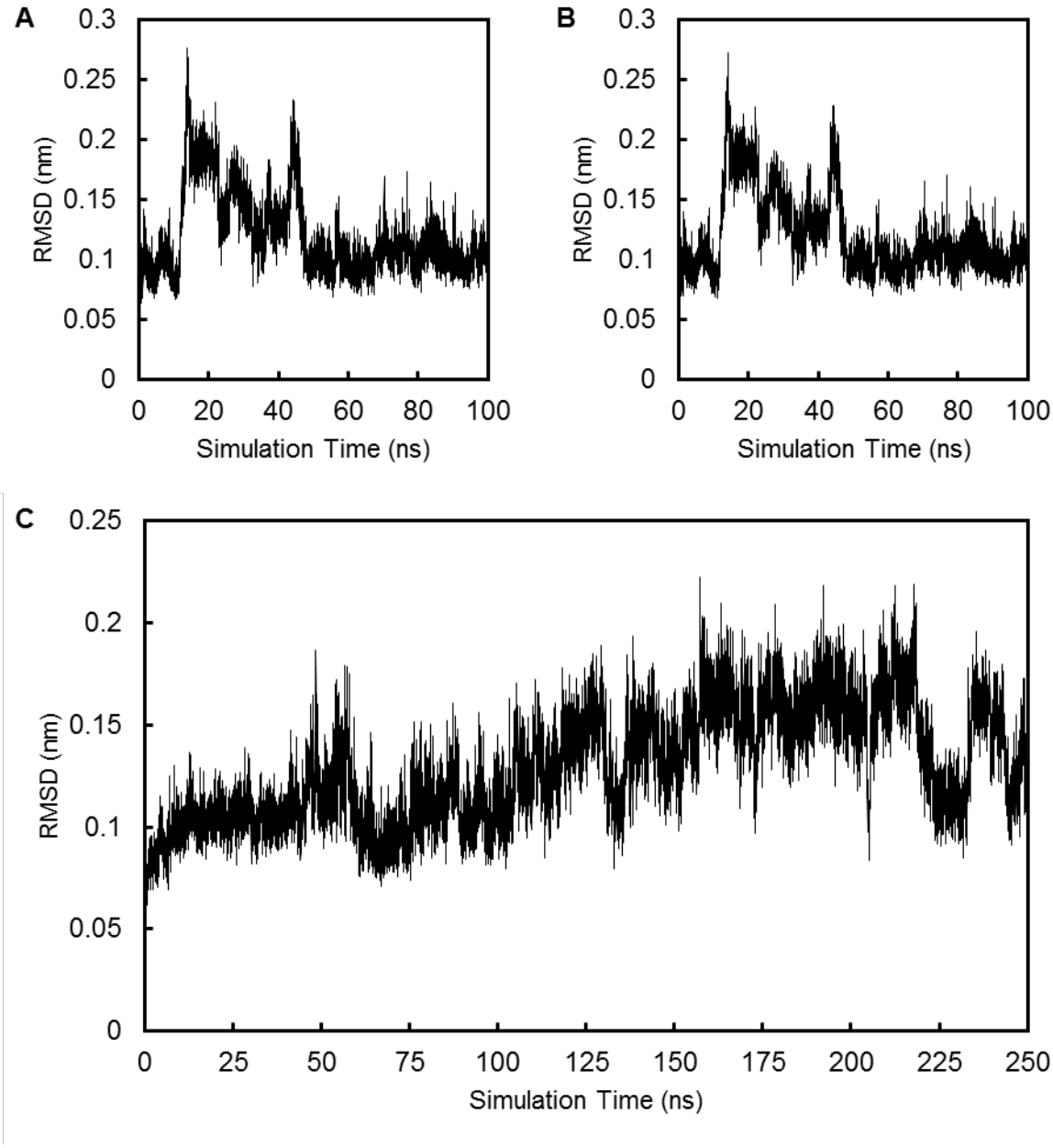


Figure S 5 | **CEACAM1 N-terminal domain molecular dynamics simulations.**

A) 100 ns production MD; plot of backbone RMSD vs energy-minimised structure. **B)** 100 ns production MD; plot of backbone RMSD vs original crystal structure. **C)** 250 ns production MD; plot of backbone RMSD vs energy-minimised structure. The maximum RMSD remained below 0.3 nm throughout all simulations indicating a stable inflexible structure.

Appendix G: CbpF YadA-like head sequence conservation

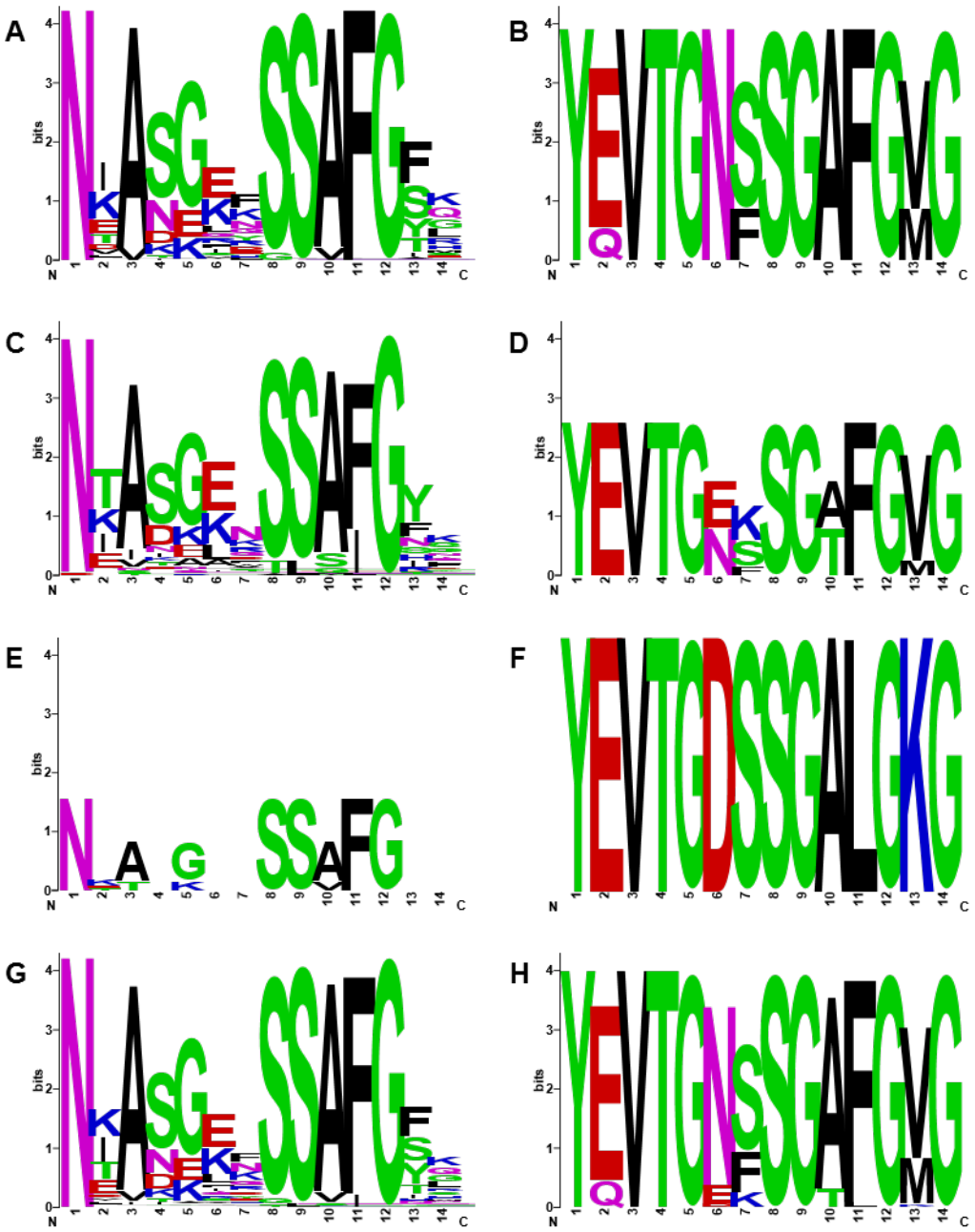


Figure S 6 | **CbpF YadA-like head sequence logos.**

Sequence logos for the different YadA-like head domains from CbpFa **(A and B)**, CbpFb **(C and D)** and CbpFc **(E and F)** were created as well as the global **(G and H)** representation. The N-terminal group of YadA-like heads were split into two groups: the final domain of the cluster **(B, D, F and H)** and all the ones prior **(A, C, E and G)**. As only one CbpFc has been identified, only one example of the final YadA-like domain exists so conservation could not be calculated **(F)**. Sequence logos were generated using WebLogo (197).

Appendix H: CbpFb Mass Spectroscopy Results

>CbpFb *[F. oralis* sp. nov. 2B3*]*
MAPAFGTGTGANSIVAGEANEATQEKSSAIGYGNKANGKFSSAFGNDNKASG
ENSSAFGRSNIASNGTSSAFGYYNTASGLRSSAFGHNNTASGENSSAFGYFN
TASEENTSAIGFKNEASGKQSSAIGYLNTASALRSSAFGINNTASGEGSSAF
GYINKVSGANSSVLGNQYEVTGNSSGAFGVGFWNSGSHLYKNEGNNSYMIGN
KNKIASGSDDNFILGNNVEIGAGVQKSVVLGDGSASGGSNTVSVGSSTLQRK
IVNVADGTISATSTDAVTGRQLYSGDGIDVNKWRTKLGVSSGGGASGGAPGD
AYTKSEADNKFTSKDDYKDANGIDVDKWKAKLGTGGGSAKHHHHHH

Accession	Description	Score	Coverage	# Unique Peptides	# PSMs	Area
NA	CbpFb	598.71	69.55	31	186	1.007E8
P00761	Trypsin OS=*Sus scrofa* PE=1 SV=1 - [TRYP_PIG]	383.05	40.69	10	101	4.929E8

Figure S 7 | **CbpFb JF1 crystal mass spectroscopy results.**

Highlighted are the regions identified within the expected sequence of the crystallised protein. The table shows a summary of the data analysis, showing no evidence for contaminating proteins except for the trypsin used to digest the protein solution. Peptide-spectrum matches (PSMs) were filtered with a 1 % false discovery rate (FDR) to remove any proteins that were only matched by a single peptide.

Appendix I: Digital Data

Common scripts used can be found at GitHub: https://github.com/mb1511

Large data sets from molecular dynamics, X-ray diffraction and whole-genome comparisons will be available by request to Prof RL Brady[1] or Dr DJ Hill[1].

[1] University of Bristol, Biomedical Sciences Building, University Walk, Bristol BS8 1TD